MethylSeqR: An R package for efficient analysis of modified nucleotide data

Hailey E. Zimmerman[1,2], Isaac Stirland[1], Ryan H. Miller[1,2], Andrew Jenkins[1,3], Erin Saito[1], Tim Jenkins[1,3,*], Jonathon T. Hill[1,3,*]
*These authors contributed equally to this work.

Affiliations
[1] Wasatch BioLabs, 2315 West 700 South, Suite R, Pleasant Grove, UT 84062 USA.
[2] Department of Biology, Brigham Young University, 701 E. University Pkwy, LSB 4102, Provo, UT 84602 USA.
[3] Department of Cell Biology & Physiology, Brigham Young University, 701 E. University Pkwy, LSB 4005, Provo, UT 84602 USA.

Email Addresses
Hailey Zimmerman: hailey@renewbt.com
Isaac Stirland: isaac@wasatchbiolabs.com
Ryan Miller: ryanhm@byu.edu
Andrew Jenkins: jenk2319@byu.edu
Erin Saito: erin@renewbt.com
Tim Jenkins: tim_jenkins@byu.edu
Jonathon T. Hill: jhill@byu.edu

Corresponding Authors
Jonathon T. Hill
Department of Cell Biology and Physiology
Brigham Young University
701 E. University Pkwy, LSB 4005
Provo, UT 84602

Tim Jenkins
Department of Cell Biology and Physiology
Brigham Young University
701 E. University Pkwy, LSB 4005
Provo, UT 84602

Keywords
DNA methylation, differential methylation analyses, Illumina methylation microarray, Oxford Nanopore sequencing, 5-methylcytosine, 5-hydroxymethylcytosine, R

Abstract

DNA methylation regulates a wide range of biological processes, including gene expression, disease progression, and cell identity. Long-read technologies now enable more comprehensive and accurate methylome analyses than ever before, but they are hindered by the computational resources needed to analyze the massive datasets. Here, we present the CH3 file format, which aids data storage and transfer by reducing file sizes by more than 95%, and the MethylSeqR R package, which builds on the CH3 format and a database backend to enable a broad range of epigenetic analyses. Together, these tools enable high-throughput methylation analysis while minimizing computational resource requirements.

Background

DNA methylation plays vital roles in regulating cellular identity and function in humans and other animals. In this process, methyl groups are added to the fifth carbon of cytosine residues, typically within cytosine-guanine dinucleotide (CpG) contexts. Aberrant methylation patterns can destabilize the genome and disrupt normal gene expression, contributing to numerous human diseases and serving as important diagnostic biomarkers[1,2]. Beyond human biology, DNA methylation offers insights into the cellular processes across numerous other animal species[3]. Accurate detection and analysis of

DNA methylation patterns are therefore essential for understanding development, disease mechanisms, and responses to environmental changes[4].

Traditional approaches to DNA methylation profiling have relied heavily on microarray technologies such as the Illumina MethylationEpic Bead Chip. These arrays, while cost-effective and high-throughput, are inherently limited in genomic coverage and resolution[5]. More recent sequencing-based methods, including bisulfite sequencing and enzymatic methyl-seq (EM-seq), have extended our ability to interrogate methylation at single-base resolution. However, they introduce additional limitations, including biases introduced during cytosine conversion and PCR amplification, as well as computational analysis challenges due to the number of mismatches introduced by the conversion process. In contrast, native methylation sequencing directly analyzes unamplified, unconverted DNA, avoiding these limitations. Platforms such as Oxford Nanopore Technologies (ONT) and PacBio have enabled native, PCR-free sequencing and subsequent genome-wide detection of DNA modifications in real time and at high resolution[6,7]. These third-generation technologies offer more comprehensive and unbiased epigenetic analyses than array-based platforms and other traditional methods[8]. These technologies present significant technical advantages in most use cases, especially when sequencing genomes from non-model organisms or exploring genomic regions that are poorly represented by commercial arrays[9]. A complete analysis of the performance of ONT sequencing is covered in the accompanying paper (cosubmission citation).

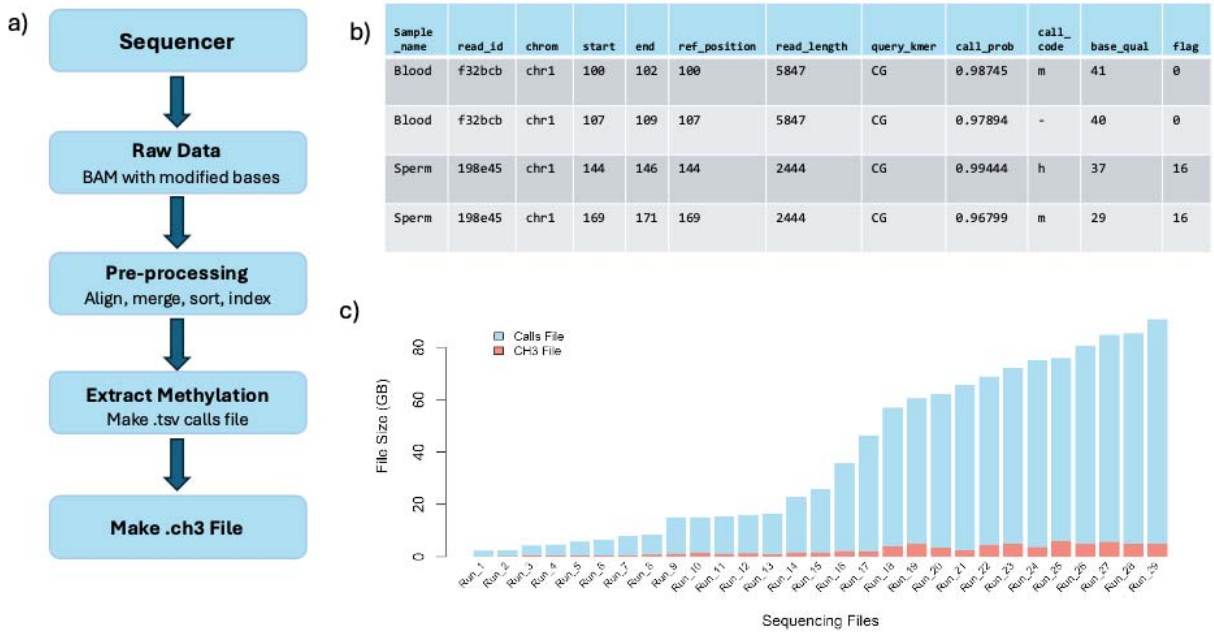
Despite its promise, native methylation sequencing data remains challenging to work with due to its large size and lack of streamlined analytical tools[10]. A single pass of the human genome (1x) produces 3-5 gigabases (Gbases) of data. Thus, sequencing at deeper coverages or with multiple samples can produce terabytes of information. Available tools, such as NanoSpring, are built to address this issue by compressing raw sequencing data[11]. However, these tools fall short in providing support for further analysis. Various software packages, such as the widely used MethylKit, provide differential methylation analysis capabilities, but were primarily developed for short-read bisulfite sequencing data[12]. Consequently, their performance is suboptimal for analyzing other data types such as nanopore sequencing[13]. As a result, there remains an urgent need for efficient tools for downstream analysis of long-read, native DNA methylation data.

To address these gaps, we developed a two-part solution: a highly-efficient modified nucleotide data storage format (CH3) that reduces nanopore methylation data size by 90-95%, and MethylSeqR, a publicly available R package for fast, flexible downstream analysis. MethylSeqR supports differential methylation analysis at single-base or regional levels, includes built-in visualization tools, and offers customizable parameters for diverse experimental designs. Here we present the package's workflow, features, and performance.

Results

CH3 - Efficient DNA Modification Data Storage

Methylation sequencing generates large volumes of raw data, typically in the form of binary alignment map (bam) files containing raw read sequences. When modified base calling is enabled, two tags are added to each sequence record in the file: one for the modified base call positions and a second with probabilities for modification at the position. In order to analyze the data, these files undergo a series of pre-processing steps, as summarized in **Figure 1a**. This process includes alignment, merging, sorting, and indexing. Finally, base modification predictions—such as methylation events—can be extracted and stored in .tsv files using tools like the modkit package (https://github.com/nanoporetech/modkit). While convenient as a human-readable format, these plain-text files are often unmanageably large, presenting challenges for downstream analysis, storage, and sharing.



Sample_name	read_id	chrom	start	end	ref_position	read_length	query_kmer	call_prob	call_code	base_qual	flag
Blood	f32bcb	chr1	100	102	100	5847	CG	0.98745	m	41	0
Blood	f32bcb	chr1	107	109	107	5847	CG	0.97894	-	40	0
Sperm	198e45	chr1	144	146	144	2444	CG	0.99444	h	37	16
Sperm	198e45	chr1	169	171	169	2444	CG	0.96799	m	29	16

Figure 1: Specification for the CH3 file format. A) The general workflow used to generate a CH3 file. B) Sample CH3 file contents. C) Memory comparison between CH3 and modkit calls formats across 29 nanopore sequencing runs. Each bar represents the total file size in gigabytes (GB) for a single sequencing run, with modkit calls files shown in light blue and the corresponding CH3 files shown in red.

To address the limitations inherent in the standard methylation output format, we developed the custom CH3 file format (named after the chemical formula for a methyl group), a compact, efficient representation of base modification calls designed for efficient compression and retrieval. This format contains all essential nucleotide modification information, including the sample name, read ID, chromosome, 0-based reference position, read length, query k-mer, call probability, call code, base quality, and

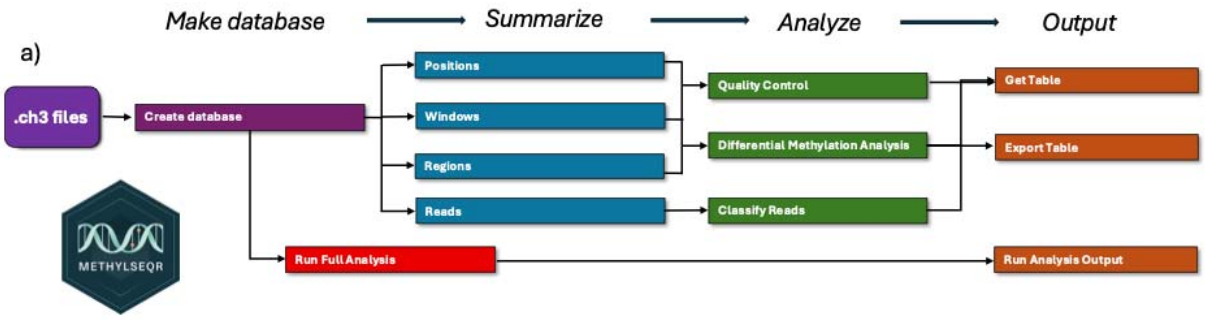
samtools flag value. The CH3 file specification is also extensible, so additional columns can be added to meet unique needs. Internally, the CH3 format leverages the Apache Parquet framework (https://parquet.apache.org/), enabling fast, column-based access and easy use in many popular programming languages such as R, Python, C++, and Java. The Parquet format reduces file size by using dictionary-based encoding for string columns, and incorporates the zstd compression algorithm (https://github.com/facebook/zstd) to efficiently represent data within the file while allowing for fast access. In addition to the inherent efficiencies of the Parquet format, the CH3 file specification also relies on the samtools flag, which encodes alignment classification (unaligned, primary, secondary, or supplementary) and strand (+ or -) into a single integer instead of having multiple columns for this information (**Figure 1b**), further reducing its size. Together, these factors reduce the size of methylation data stored in TSV files to just 5–10% of their original size, ensuring that large-scale epigenomic datasets can be stored, transferred, and queried using minimal computational resources (**Figure 1c**).

MethylSeqR - An R Package for Analyzing Modified DNA

Current R packages for analyzing modified DNA have two major limitations. First, they are built around a bisulfite conversion paradigm. Thus, they do not support direct detection of modifications and are generally limited to methylation calls. However, third-generation sequencing platforms, such as the Nanopore platform, can directly detect a large number of modifications. Second, the amount of data generated from sequencing is extremely large. For example, there are approximately 30 million CpG dinucleotides in

the human genome. At a sequencing depth of 10x, that is 300 million calls per sample. This is orders of magnitude larger than the 900,000 data points generated by the Epic Array and large enough to use the entire vector memory limit on machines with modest hardware specifications.

To address both of these issues, we have created the MethylSeqR package (outlined in **Figure 2a**). This package uses a DuckDB database backend, allowing users to conduct large, complex analyses on a wide range of hardware[14]. The database is managed internally to maintain ease of use and familiarity for users accustomed to using R's conventional data structures and paradigms (e.g. piping) by managing the database internally. The package also provides several built-in functions for working with the base call data to facilitate common workflows, including QC, data summarization by position, read, window, and region, and differential methylation analysis. An example pipeline with explanations for each function is shown in **Figure 2b-c**. Below, we walk through the major steps of an analysis workflow and highlight package functionality and outputs.



b)

```
ch3_db <- make_ch3_db(
            ch3_files = "/desktop/ch3_files",
            db_name = "my_data") |>
  summarize_ch3_windows() |>
  calc_ch3_diff(
            call_type = "windows",
            cases = c("sperm"),
            controls = c("blood")) |>
  collapse_ch3_windows() |>
  export_ch3_table(
            table ="mod_diff_windows",
            out_path = "/desktop/results")
```

c)

Function	Required Arguments	Description
make_ch3_db	ch3_files db_name	Processes a directory of ch3 files and creates a database containing filtered and structured methylation call data.
summarize_ch3_windows	ch3_db	Summarizes methylation data by creating sliding windows over the genome
calc_ch3_diff	ch3_db cases controls	Calculates differential methylation between specified case and control groups
collapse_ch3_windows	ch3_db	Identifies and merges nearby significantly differentially methylated windows after a differential methylation analysis is called

Figure 2: MethylSeqR Core Design and Workflow. A) Schematic of possible MethylSeqR workflows, starting from the input of CH3 files and progressing through database creation, summarization, analysis, and output. B) Example pipeline for a full differential methylation pipeline using R's native pipe syntax. Pipeline begins with creating a database, then summarizes data using sliding windows, and performs a differential methylation analysis. Finally, significant windows are collapsed into merged regions, and the results are exported to a user-defined path. C) The core functions used in the example pipeline in Panel B, along with their required arguments and a brief description.

Database Creation and Filtering

The first step in any MethylseqR pipeline is to create a database with all samples to be included in the analysis (**Figure 2a**). The CH3 file format is designed to store the data from a single sample, while the database combines multiple samples into a single data structure, a calls table, for analysis (**Supplemental Figure 1**). The database is written into the local directory as “filename.ch3.db”, and is updated as tables are created within it, allowing persistent storage of intermediate data structures and limiting the need to rerun calculations between sessions or to maintain large amounts of data in memory during multistep pipelines.

During this stage, the data can also be filtered for a variety of QC parameters, including minimum read length, minimum call probability, and minimum base quality. Additionally, users can choose to focus on specific types of modifications or combinations of several modifications. For example, users importing data from CH3 files containing methylation and hydroxymethylation can choose to use either modification, combine both into a single call, or retain all modifications. This level of customization empowers researchers to refine their workflows and explore epigenetic patterns in a way that aligns with their biological questions and data characteristics.

Summarizing DNA methylation information over positions, pre-defined regions, or tiling windows

The modification calls stored in the database can be summarized in many ways (**Figure 2a**). MethylSeqR is able to summarize DNA methylation data across individual genomic positions, predefined regions (such as promoters, CpG islands, or introns), or sliding windows for further analysis. It also supports read-level summarization for mixture analysis (described below). These summaries can then be used for downstream analyses of these unique sites of interest. For more detailed explanations of all functions within the package, refer to the MethylSeqR vignette.

Users have the accessibility to export any resulting summary tables directly into their R environment for further exploration with the get_table() function. This includes data summarized by positions, windows, regions, reads, or a differential methylation analysis. Example summarization table outputs using a synthetic dataset are shown in **Supplemental Figure 2**. Additionally, users may save the data table as a CSV file for external use with export_table().

Differential Methylation Analysis

A differential methylation analysis can be performed to identify statistically significant differences in DNA methylation levels at individual bases or genomic windows between two different sample groups. This analysis is essential for detecting epigenetic changes that may play roles in gene regulation, development, or disease

processes such as cancer and neurological disorders[15]. MethylSeqR provides multiple methods for conducting differential methylation analysis, including logistic regression, Fisher's exact test, and a rapid implementation of Fisher's test (r_fisher). MethylSeqR generates differential statistics for each group, including modification fraction difference and p-value calculations for regions. For an example output table, see **Supplemental Figure 3.** The output of a differential methylation analysis can be plotted as a volcano plot (**Figure 3a**) or a methylation correlation plot (**Figure 3b**).

To test the accuracy of the statistical analyses, we created a synthetic dataset with known modification differences between example case and control groups. When differential analysis was performed using MethylSeqR, our function accurately identified all positions and their corresponding modification difference between the groups.

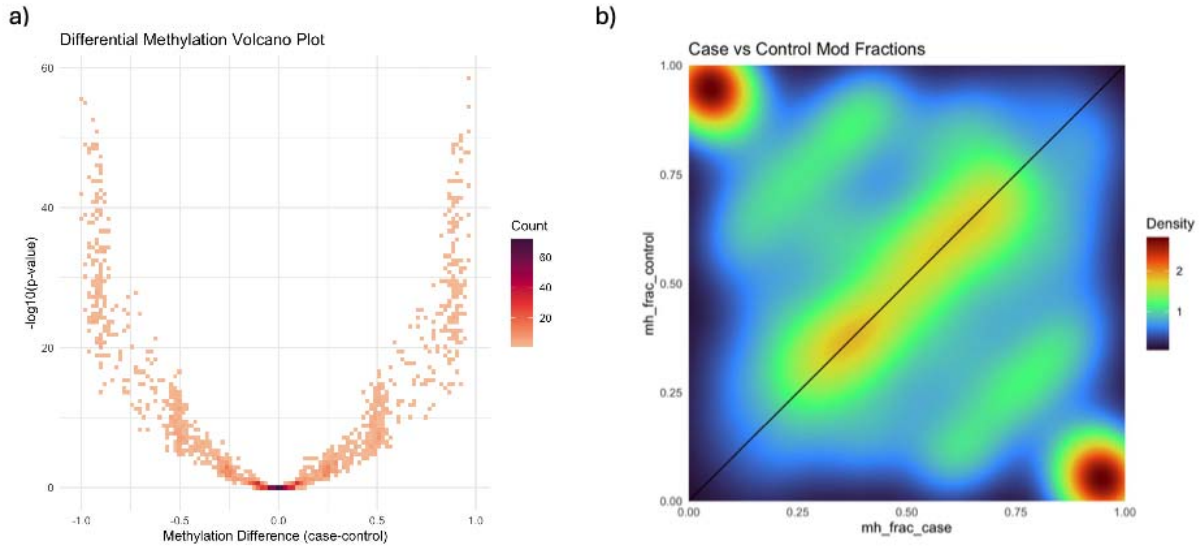


Figure 3: Example Output Plots for Differential Expression Analysis. A) Example differential volcano plot generated with a custom MethylSeqR function. Each square represents a genomic window/position. The x-axis is the methylation-fraction difference between the case and control samples; the y-axis is the statistical significance. Density correlates with darker colors, and points near the center indicate regions with little or no difference. B) Case-vs-control fraction concordance plot built with a custom function. Each location's modified fraction in the case (x-axis) is plotted against the corresponding fraction in the control (y-axis). The background is a 2-D kernel-density surface: red/yellow areas mark high point density, blue areas low density.

QC metrics: descriptive statistics, sample correlation, and PCA

In addition to its core analysis pipelines, MethylSeqR provides a suite of tools for whole methylome characterization, including general quality control metrics, sample-level comparisons, and dimensionality reduction. Users can visually assess key characteristics of their methylation data using built-in plotting functions (**Figure 4**). Two primary visualization options are available: read coverage per base, which provides insight into sequencing depth and data reliability, and percent methylation per base, which reflects the underlying biological signal. These metrics are stored in the package's core data structures and are easily accessible for customized visualizations.

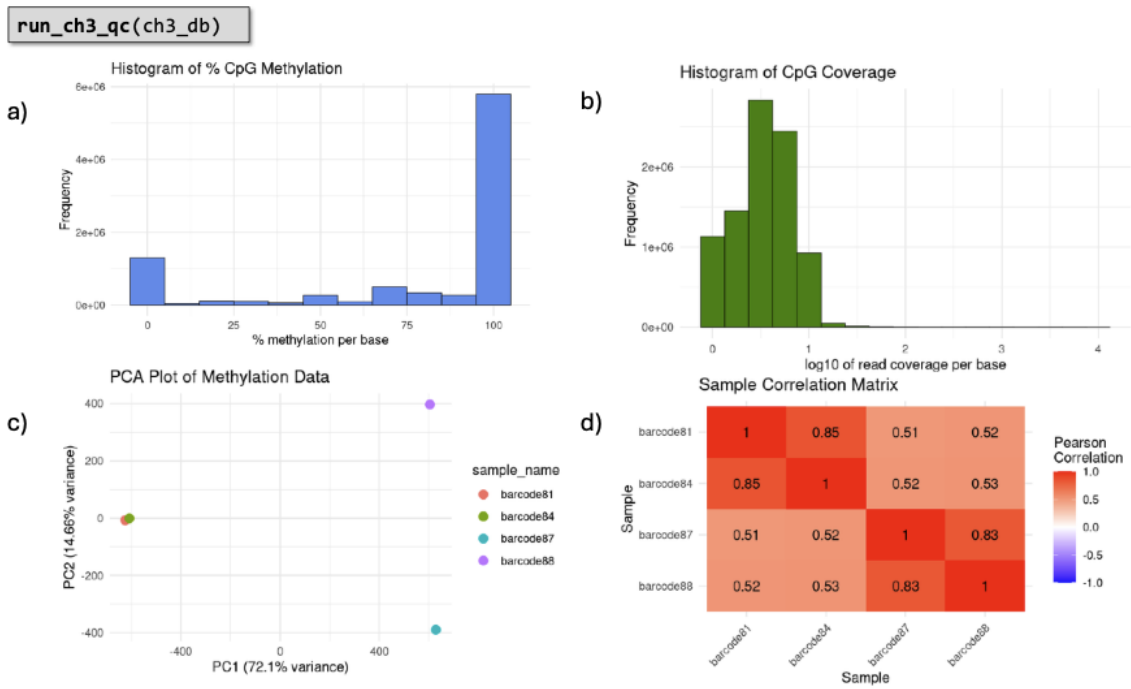


Figure 4: Descriptive Statistics and Quality Control Function Outputs. A) Percent methylation per base graph: This histogram shows the distribution of CpG methylation percentages across all bases for a sample. It provides an overview of the methylation landscape, allowing users to quickly assess whether methylation values are skewed, bimodal, or evenly distributed. B) Read coverage graph: This histogram displays the distribution of CpG read coverage (on a log10 scale) per base. It helps evaluate the depth and consistency of sequencing coverage, which is critical for assessing the reliability of methylation calls. C) PCA plot: The principal component analysis (PCA) plot reduces high-dimensional methylation data into two main components (PC1 and PC2) that capture the largest variance. This visualization helps identify clustering of samples and potential batch effects or outliers. D) Correlation plot: The sample correlation matrix helps visualize the pairwise Pearson correlation coefficients between samples based on

methylation values. It quantifies the linear relationship between samples, helping assess similarity and reproducibility across experimental conditions.

To compare methylation profiles across samples, MethylSeqR includes correlation plots based on the Pearson correlation coefficient, a measure of the linear relationship between two continuous variables. Additionally, a principal component analysis (PCA) function is provided to explore global patterns in methylation data and detect clustering among samples. PCA reduces the dimensionality of high-throughput data by projecting it into a smaller set of orthogonal components that capture the greatest variance, helping reveal biological trends or batch effects across the dataset.

Performance

MethylSeqR is designed to be computationally efficient and capable of analyzing large datasets in practical runtimes. Thus, we tested the performance of several key analysis steps on various dataset sizes. The most computationally intense step within the analysis is summarization. The database is initially created with one row per methylation call. These calls are then summarized into positions, reads, regions, or sliding windows across all samples, which requires large amounts of database read and write calls. As shown in **Figure 5a**, the time to conduct the summarization depends on both the type of summarization and the size of the dataset. Importantly, however, the computation completes for all dataset sizes, even when larger than the memory available. Differential analysis also shows the same smooth increase in time without failing (**Figure 5b**).

Finally, we compared run times between an analysis server with 250 GB of RAM and 96 cpu cores compared to a personal computer with 16 GB of RAM and 8 cores (**Figure 5c**). While run times were significantly longer on the personal computer, times were still reasonable, and all analyses were completed.

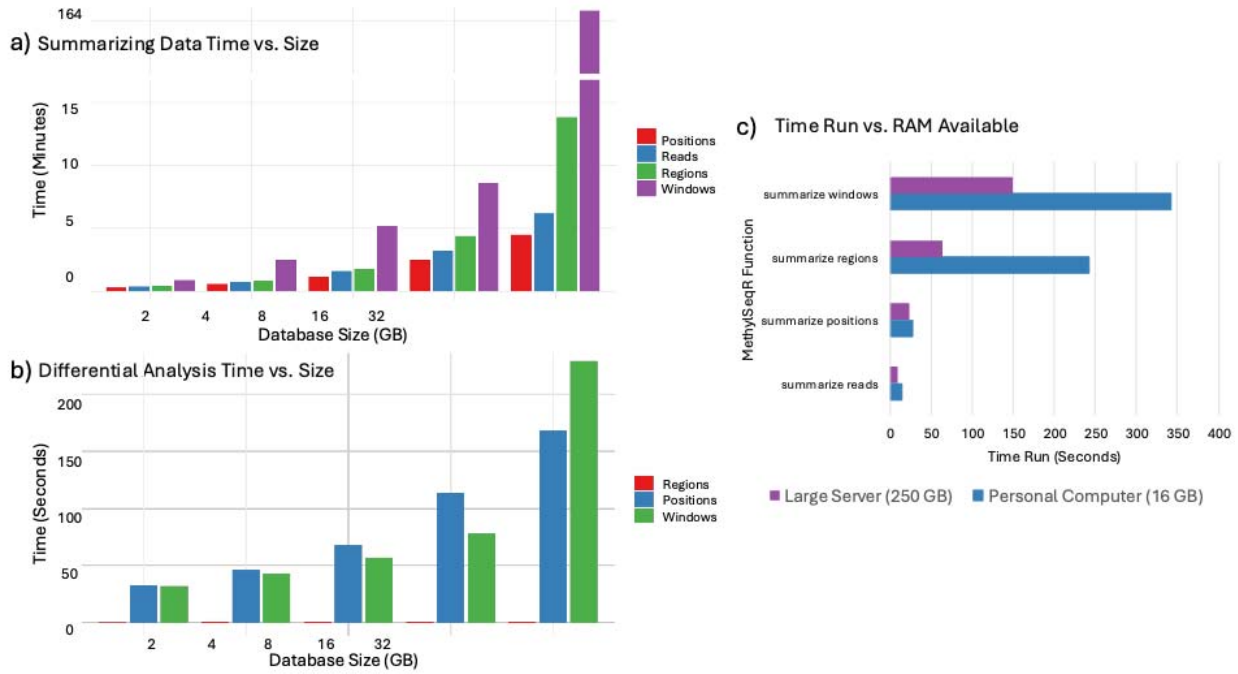


Figure 5: MethylSeqR Performance on different hardware configurations. A) Summarizing Data Time vs. Size for different data collection functions. This includes summarizing methylation data by windows (100 bp step size), reads, genomic positions, or regions. B) Differential Analysis Time vs. Size. This figure depicts a differential analysis between three available options (regions, positions, and windows). These functions were run on a personal computer with 16GB of RAM. C) Time vs. RAM available for 4 functions on two different machines - one is a personal computer with

16GB of RAM available, while the other is a large server with 250GB of memory available. The dataset consisted of 4 samples of 5x coverage.

Discussion

Analyses of DNA methylation and other nucleotide modifications have enormous potential to provide novel biological insights into both natural processes and disease mechanisms. While such analyses have been conducted for more than a decade, the recent development of direct modified base calling on the NanoPore and PacBio platforms has greatly increased genomic coverage of these assays. For example, the current Epic Array has approximately 900 thousand of the approximately 30 million CpG sites in the genome (approximately 3%). However, sequencing regularly provides information on over 28 million sites (93%) (see cosubmission). This is a significant leap forward that will provide significant insights missed by previous technologies.

Although the increased genomic coverage provided by third-generation sequencing is very welcome, it also results in a significant increase in the size of the data files produced, hindering analysis by requiring large amounts of computational resources and specialized bioinformatics know-how to manage the data. Here, we have presented two tools to address these issues: the CH3 file format and the MethylSeqR R package. These tools, used together, will open up modified nucleotide analysis to more users and drive wider adoption across the field.

One of the primary challenges with large datasets lies in the cost and complexity of storing and transferring them—difficulties that will only intensify as data generation continues to grow. Thus, we created a file format specification, CH3, that addresses these problems by vastly decreasing the file size for modified base data. Analyses presented here show that CH3 typically reduces file sizes by 95%. These reductions come from a number of sources, including the removal of redundant and unnecessary data fields, efficient compression algorithms, and careful selection of data types. File access and read times are also improved by using column-based storage and internal indexing. Importantly, the file specification is designed to be used with any modification code and allows for additional columns to be added by the user to address specific use cases and to future-proof the format. Thus, this file format will provide a strong foundation for future studies on nucleotide modifications.

We also present in this paper the first R package designed specifically to work with native modified base calling. As the first tool, we do not have any other tools to compare to, but we were able to establish that the tool is able to handle large amounts of data and provide accurate results using a synthetic dataset. We have also incorporated a large number of functions to access and interact with the data, measure QC, and conduct differential methylation analyses on positions, sliding windows, annotated regions, or reads. As we look to the future, our team plans to further enhance MethylSeqR's analytical capabilities by integrating sophisticated machine learning algorithms, such as

k-Nearest Neighbors (KNN) and regularized regression models, further advancing modified base analysis in a number of use cases[16].

By dramatically reducing data storage requirements and offering an accessible, purpose-built R package, our work lowers key technical barriers that have limited the broader adoption of third-generation sequencing for epigenomic research. As the field continues to evolve, we anticipate that these tools will not only accelerate discovery but also inspire the development of new methodologies for understanding the biological roles of modified nucleotides in health and disease.

Conclusions

The CH3 file format and MethylSeqR R package introduced in this study represent important advances for the analysis of native DNA modifications using third-generation sequencing. By addressing key challenges in data size, accessibility, and analytical flexibility, these tools enable researchers to efficiently store, manage, and analyze high-throughput methylation data across a range of hardware configurations. The CH3 file format significantly reduces storage demands while preserving analytical utility, and MethylSeqR provides a comprehensive and user-friendly R-based framework for differential methylation analysis, quality control, and visualization. Together, they remove critical technical barriers and empower broader use of native nucleotide modification data across biological research. As interest in epigenetic mechanisms continues to expand, these tools will serve as a foundational platform for scalable,

accessible, and reproducible studies of DNA methylation and other base modifications in both health and disease contexts.

Methods

CH3 File Specification

To address the challenges of storing large-scale native base modification data, we developed the CH3 file format—a compressed, columnar format based on Apache Parquet. CH3 includes required fields such as read ID, chromosome, position, modification type, and probability, and supports additional optional metadata. File reduction is achieved via dictionary encoding and Zstandard compression, both natively implemented as part of the Parquet standard. The format also supports fast, column-based access and is compatible with MethylSeqR and common data science tools (e.g., DuckDB, PyArrow). Full specification details, including field definitions and encoding rules, are available at https://github.com/Wasatch-Biolabs-Bfx/CH3.

MethylSeqR Implementation

MethylSeqR is implemented in R and SQL and structured around modular functions that facilitate scalable, memory-efficient analysis of native base modification data. At its core, the package uses a DuckDB backend to support disk-based storage and efficient SQL-like querying of CH3-formatted data. Key functions are organized into workflow stages, including database creation (create_ch3_database()), read filtering (filter_ch3_reads()),

summarization (summarize_ch3_positions(), summarize_ch3_windows(), summarize_ch3_regions()), and differential methylation analysis (calc_ch3_diff()) using either Fisher's exact test or logistic regression. Additional tools provide sample correlation (calc_ch3_samplecor()), read classification (classify_ch3_reads()), quality control, and dimensionality reduction via PCA. Each function is optimized to perform reliably on systems with limited RAM and supports both single-sample and multi-sample analysis workflows. Extensive helper functions manage database connections, cleanup, and internal metadata. MethylSeqR supports piping syntax and outputs tidy data frames, making it highly compatible with the broader tidyverse ecosystem. Full source code and documentation are available at: https://github.com/Wasatch-Biolabs-Bfx/MethylSeqR.

Testing Data

We created a synthetic dataset to test the accuracy of our summarization functions and differential analysis. We hope this will also enable a fast workflow through all steps of the package. The positions are divided into deliberately chosen effect-size bands: eight groups with absolute methylation-fraction shifts of ±0.90, ±0.50, ±0.25 and ±0.10 (60 loci per band), and two mid-fraction bands (0.25–0.50 and 0.50–0.75; 120 loci each). We also included randomized noise for more realistic visuals. This dataset contains two CH3 files and is also publicly available for users to create a database and test the package under the inst/extdata/ directory in the GitHub repository.

Performance Benchmarking

Performance testing was conducted on two markedly different hardware configurations to assess scalability and resource utilization. The first environment was a MacBook Pro equipped with an M3 processor and 16 GB of RAM, representing a typical end-user system. The second environment was a high-performance server with 250 GB of available memory and an AMD Threadripper processor with 96 cores, providing substantially greater computational capacity.

Declarations

Software Availability and Requirements:

Project name: MethylSeqR

Project home page: https://github.com/Wasatch-Biolabs-Bfx/MethylSeqR

Operating system(s): Platform independent (works on Windows, macOS, and Linux wherever R is available).

Programming language: R

Other requirements: R package dependencies include: arrow, DBI, dplyr, dbplyr, duckdb, duckplyr, ggplot2, progress, tidyr, utils, readr (Imports). Optional for vignettes/docs: knitr, rmarkdown (Suggests).

License: Personal and Internal Research License (v1.1) Based on PolyForm Noncommercial License 1.0.0, with an Additional Grant of Rights (Internal Research Use

Exception) provided for internal research use by commercial entities.

Any restrictions to use by non-academics: Internal evaluation and research only.

Funding

No funding

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The dataset analyzed during the current study is available in the MethylSeqR repository through the Wasatch-Biolabs-Bfx organization, at https://github.com/Wasatch-Biolabs-Bfx/MethylSeqR.

Competing interests

JH and TJ are cofounders of Wasatch Biolabs, which is currently applying for patents for the CH3 file format and MethylSeqR package.

Authors' contributions

HZ and JH wrote the code for the MethylSeqR package, as well as writing this manuscript. RM and AJ assisted in the development of the package and writing of this paper. IS assisted in writing and figure generation. ES assisted in writing and editing the manuscript. TJ helped oversee the project and provided guidance for methylation analysis standards.

References

1. Locke WJ, Guanzon D, Ma C, et al. DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Front Genet*. 2019;10:1150. doi:10.3389/fgene.2019.01150

2. Ahsan M, Ek WE, Rask-Andersen M, et al. The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. Greally JM, ed. *PLOS Genet*. 2017;13(9):e1007005. doi:10.1371/journal.pgen.1007005

3. Klughammer J, Romanovskaia D, Nemc A, et al. Comparative analysis of genome-scale, base-resolution DNA methylation profiles across 580 animal species. *Nat Commun*. 2023;14(1):232. doi:10.1038/s41467-022-34828-y

4. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet*. 2013;14(3):204-220. doi:10.1038/nrg3354

5. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1):208. doi:10.1186/s13059-016-1066-1

6. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. 2017;14(4):407-410. doi:10.1038/nmeth.4184

7. Flusberg BA, Webster DR, Lee JH, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*. 2010;7(6):461-465. doi:10.1038/nmeth.1459

8. Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life*. 2021;12(1):30. doi:10.3390/life12010030

9. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020;21(10):597-614. doi:10.1038/s41576-020-0236-x

10. Fu Y, Timp W, Sedlazeck FJ. Computational analysis of DNA methylation from long-read sequencing. *Nat Rev Genet*. Published online March 28, 2025. doi:10.1038/s41576-025-00822-5

11. Meng Q, Chandak S, Zhu Y, Weissman T. Reference-free lossless compression of nanopore sequencing reads using an approximate assembly approach. *Sci Rep*. 2023;13(1):2082. doi:10.1038/s41598-023-29267-8

12. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 2012;13(10):R87. doi:10.1186/gb-2012-13-10-r87

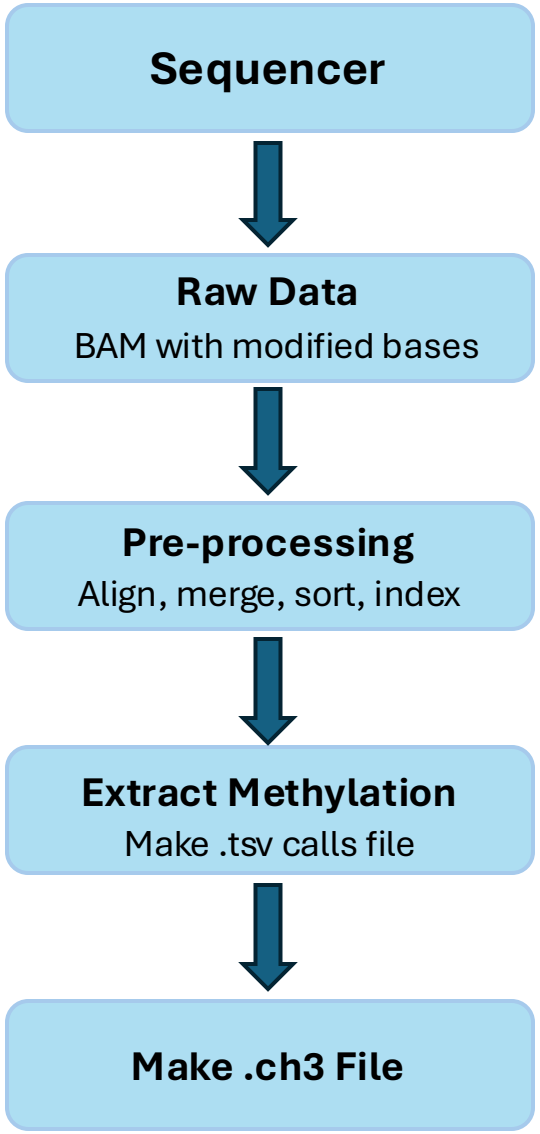
13. Meyer D, Barth E, Wiehle L, Marz M. diffONT: predicting methylation-specific PCR biomarkers based on nanopore sequencing data for clinical application. Published online February 20, 2025. doi:10.1101/2025.02.17.638597

14. Raasveldt M, Mühleisen H. DuckDB: an Embeddable Analytical Database. In: *Proceedings of the 2019 International Conference on Management of Data*. ACM; 2019:1981-1984. doi:10.1145/3299869.3320212

15. Li L, Chen R, Zhang H, et al. The epigenetic modification of DNA methylation in neurological diseases. *Front Immunol*. 2024;15:1401962. doi:10.3389/fimmu.2024.1401962

16. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1-22.

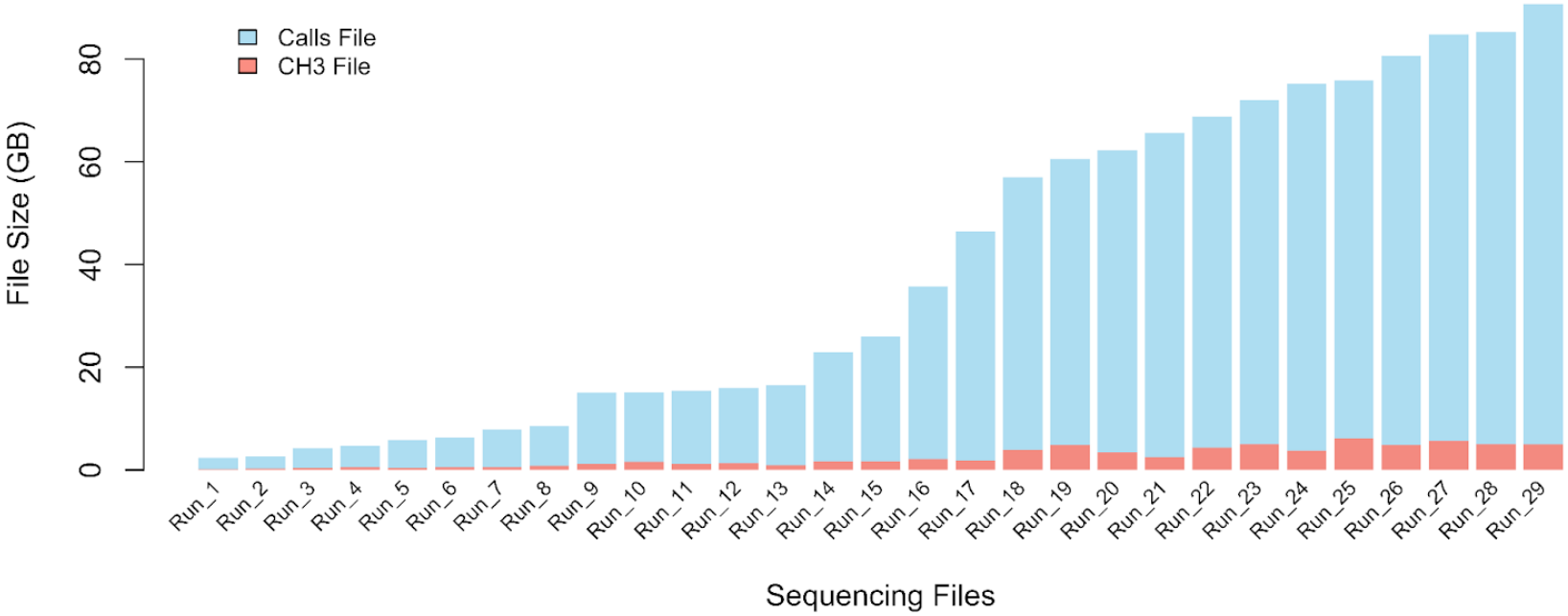
a)



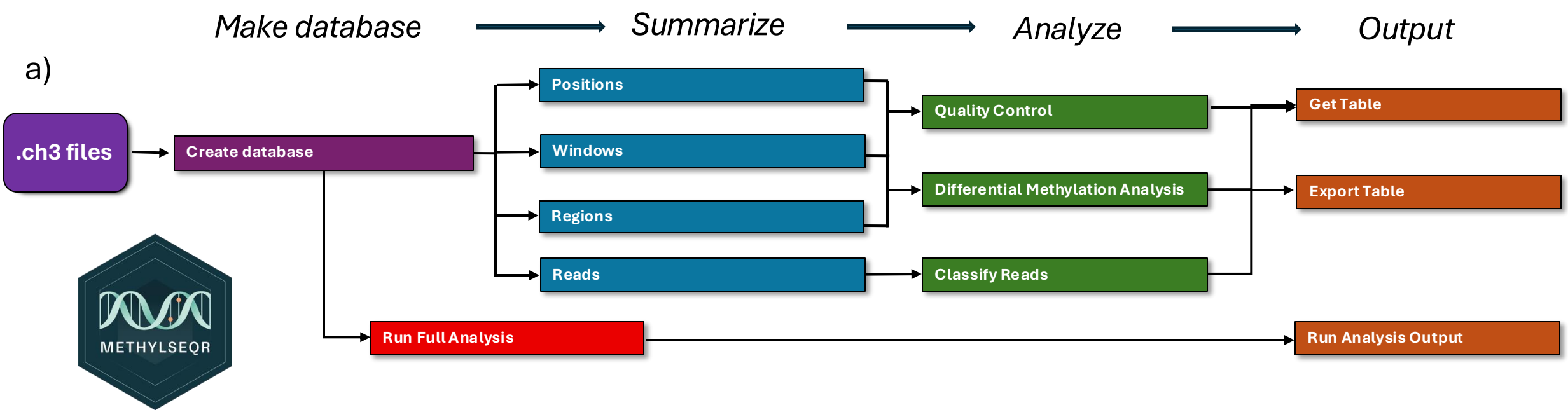
b)

Sample _name	read_id	chrom	start	end	ref_position	read_length	query_kmer	call_prob	call_ code	base_qual	flag
Blood	f32bcb	chr1	100	102	100	5847	CG	0.98745	m	41	0
Blood	f32bcb	chr1	107	109	107	5847	CG	0.97894	-	40	0
Sperm	198e45	chr1	144	146	144	2444	CG	0.99444	h	37	16
Sperm	198e45	chr1	169	171	169	2444	CG	0.96799	m	29	16

c)



a)



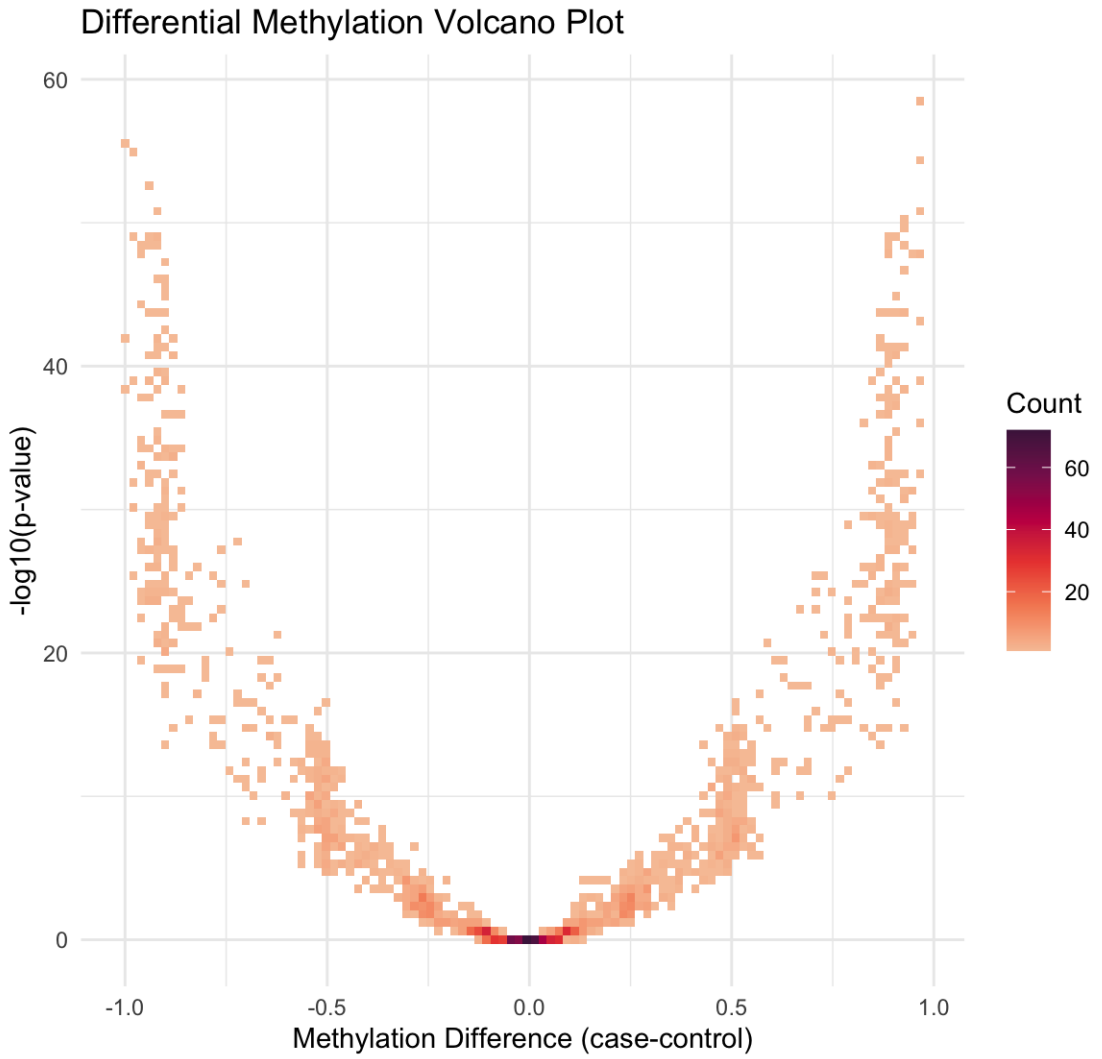
b)

```
ch3_db <- make_ch3_db(
            ch3_files = "/desktop/ch3_files",
            db_name = "my_data") |>
  summarize_ch3_windows() |>
  calc_ch3_diff(
            call_type = "windows",
            cases = c("sperm"),
            controls = c("blood")) |>
  collapse_ch3_windows() |>
  export_ch3_table(
            table ="mod_diff_windows",
            out_path = "/desktop/results")
```

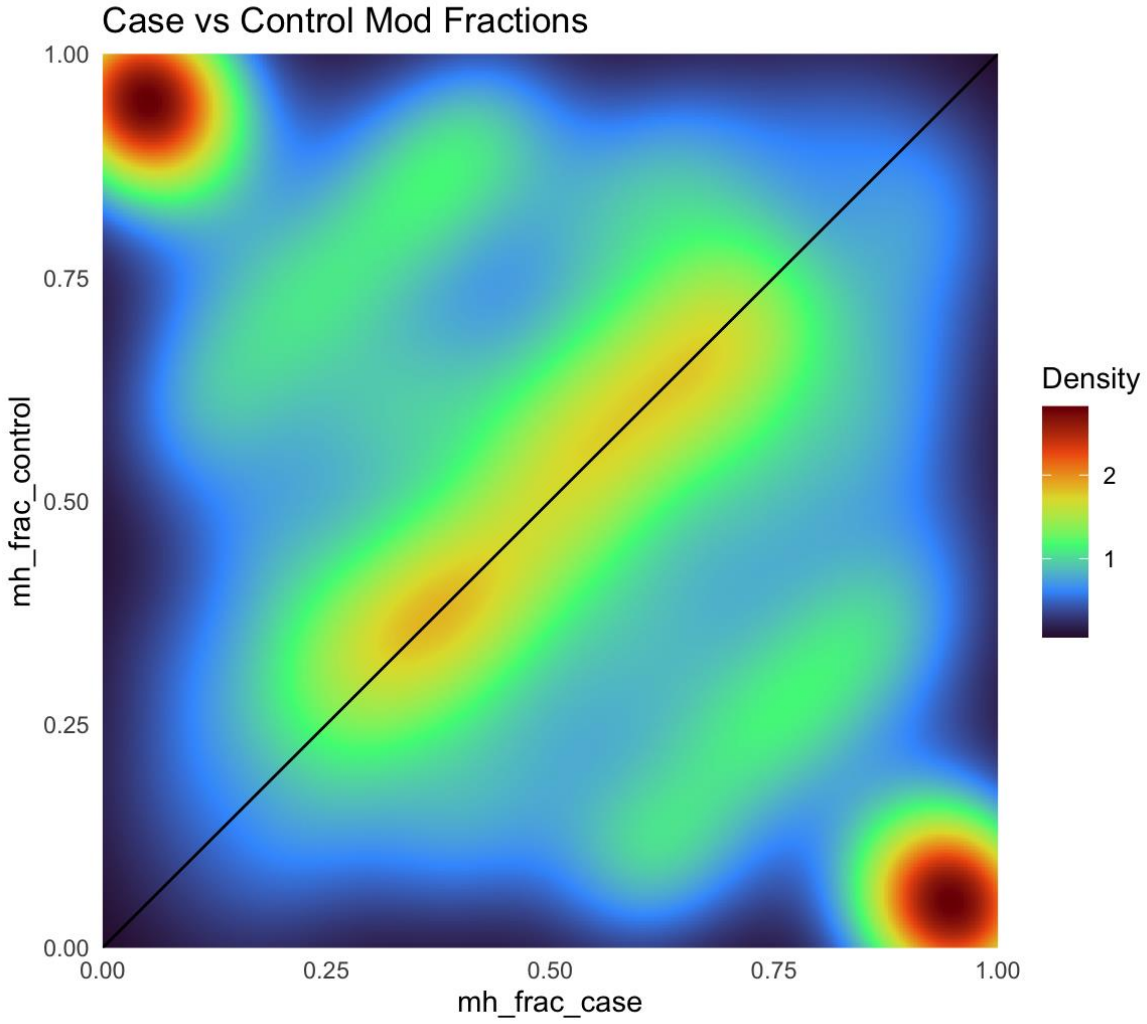
c)

Function	Required Arguments	Description
make_ch3_db	ch3_files db_name	Processes a directory of ch3 files and creates a database containing filtered and structured methylation call data.
summarize_ch3_windows	ch3_db	Summarizes methylation data by creating sliding windows over the genome
calc_ch3_diff	ch3_db cases controls	Calculates differential methylation between specified case and control groups
collapse_ch3_windows	ch3_db	Identifies and merges nearby significantly differentially methylated windows after a differential methylation analysis is called

a)

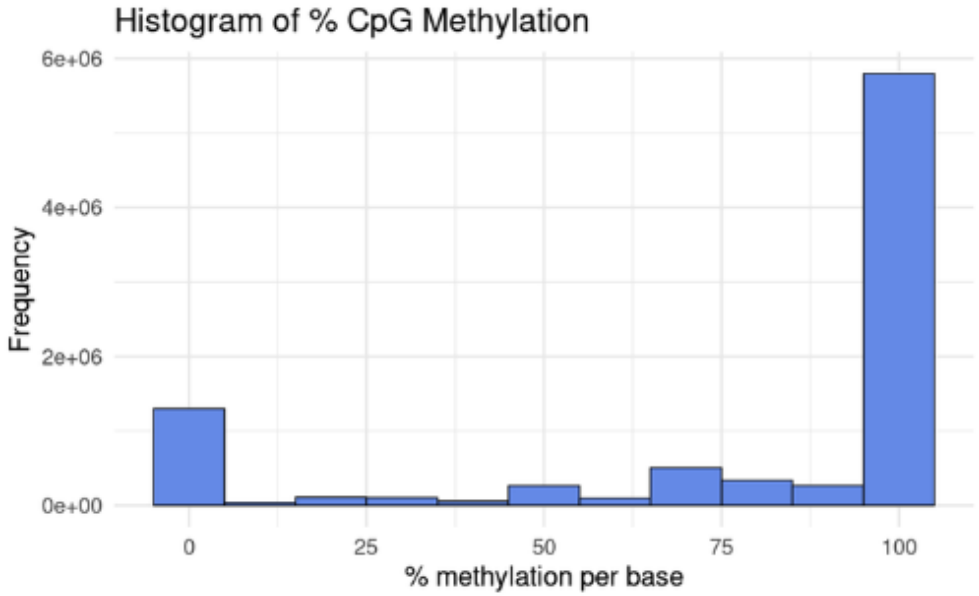

Differential Methylation Volcano Plot
60
40
20
0
-log10(p-value)
-1.0
-0.5
0.0
0.5
1.0
Methylation Difference (case-control)
Count
60
40
20

b)

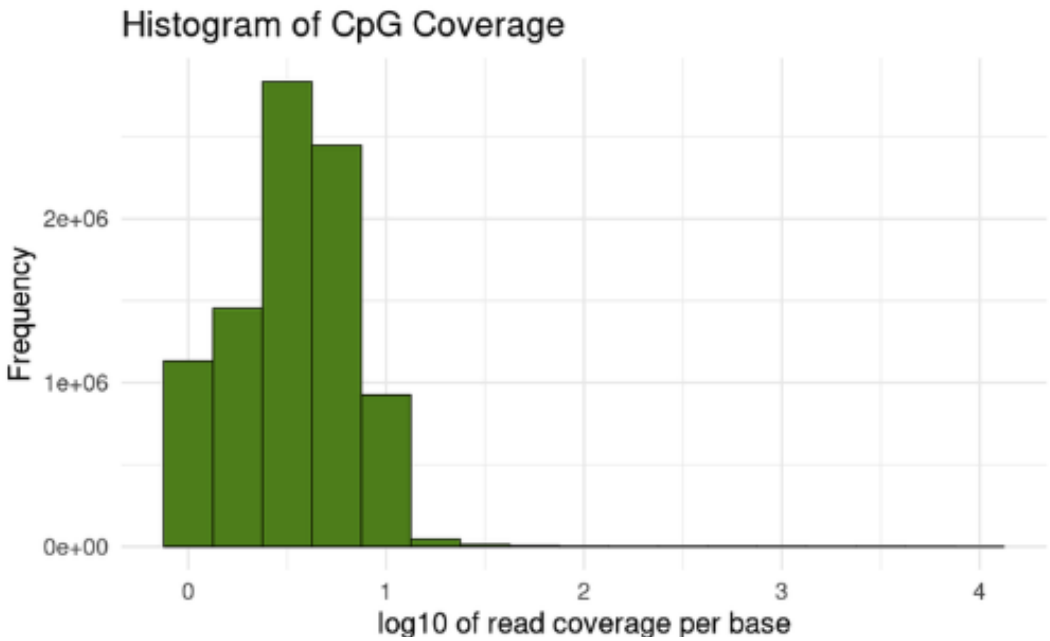

Case vs Control Mod Fractions
1.00
0.75
0.50
0.25
0.00
mh_frac_control
0.00
0.25
0.50
0.75
1.00
mh_frac_case
Density
2
1

```
run_ch3_qc(ch3_db)
```

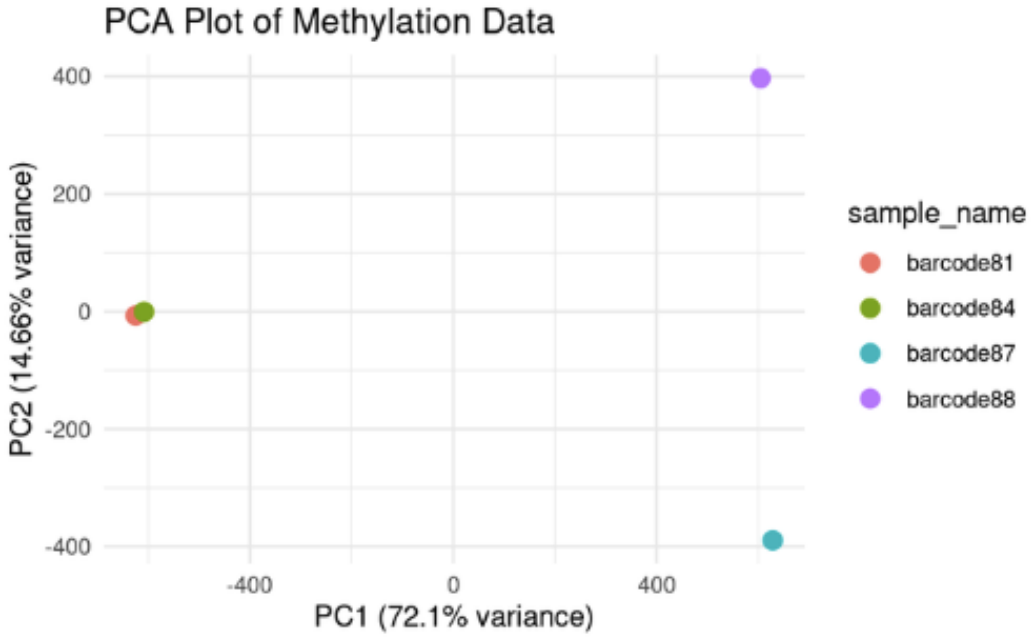
a)



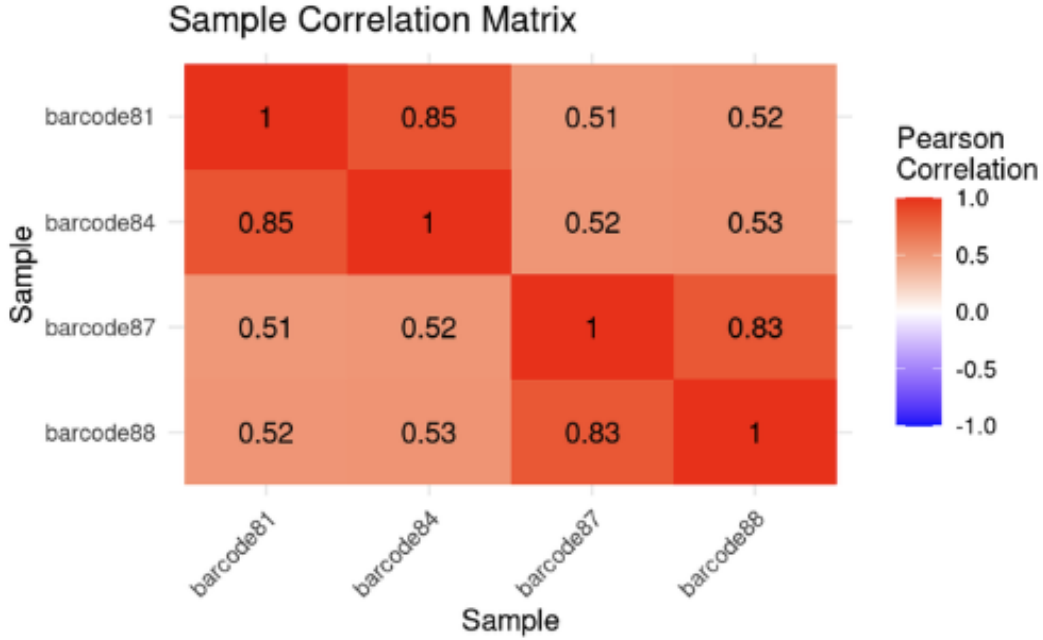
b)

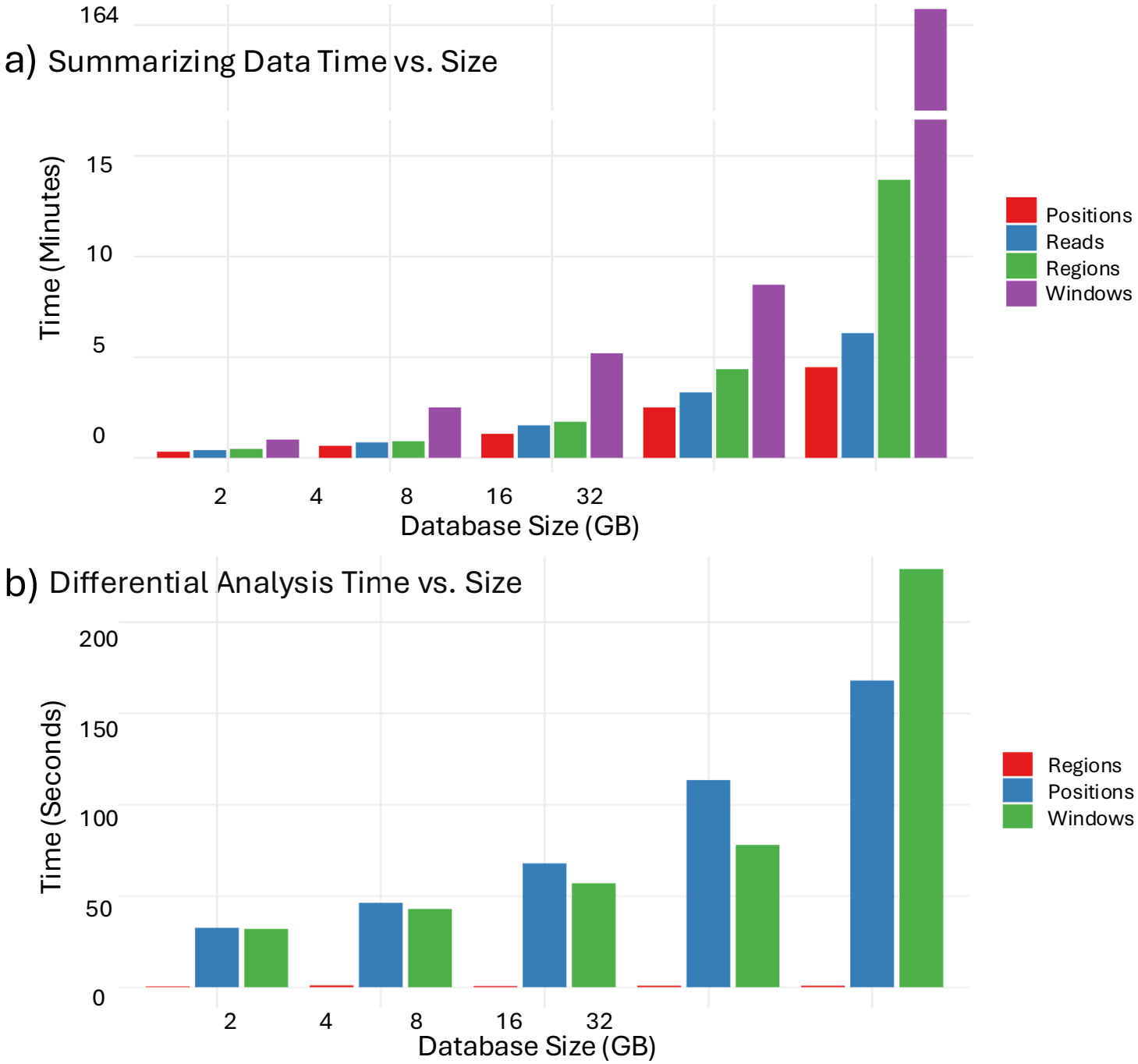


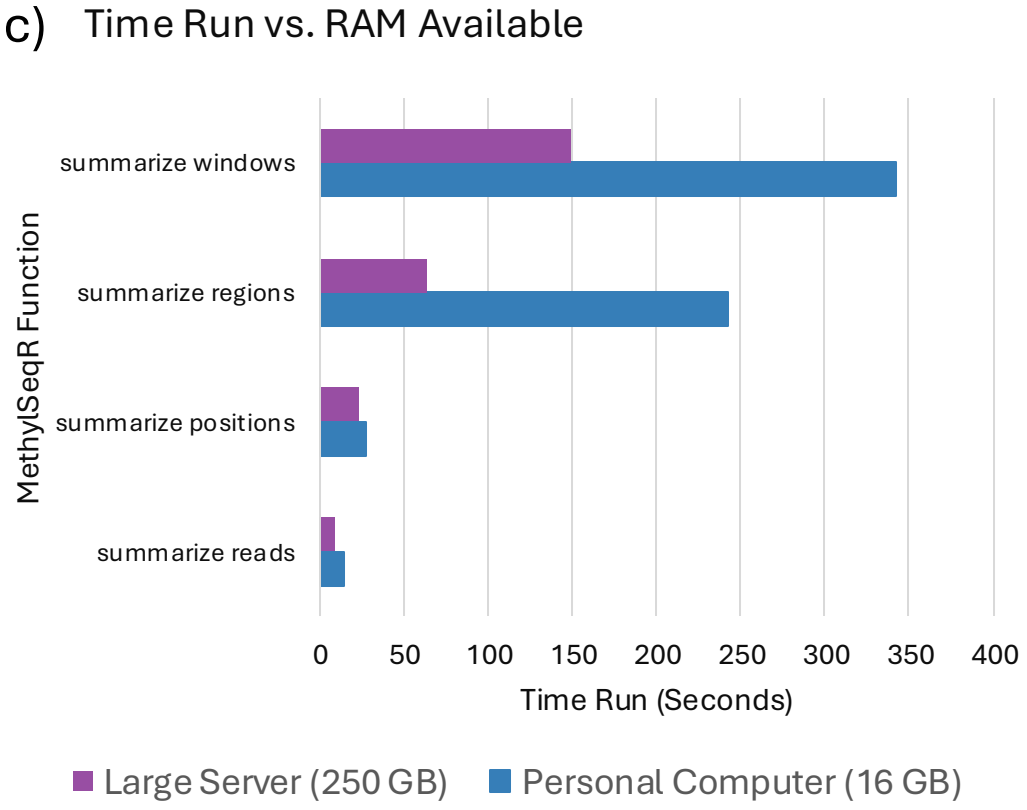
c)



d)




a) Summarizing Data Time vs. Size
164
15
10
5
0
Time (Minutes)
2
4
8
16
32
Database Size (GB)
Positions
Reads
Regions
Windows
b) Differential Analysis Time vs. Size
200
150
100
50
0
Time (Seconds)
2
4
8
16
32
Database Size (GB)
Regions
Positions
Windows


c) Time Run vs. RAM Available
summarize windows
summarize regions
summarize positions
summarize reads
MethylSeqR Function
0
50
100
150
200
250
300
350
400
Time Run (Seconds)
Large Server (250 GB)
Personal Computer (16 GB)