Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cancer Res. Author manuscript; available in PMC 2018 November 01.

Published in final edited form as:

Cancer Res. 2017 November 01; 77(21): e27–e30. doi:10.1158/0008-5472.CAN-17-0330.

Integrating DNA methylation and hydroxymethylation data with the mint pipeline

Raymond G. Cavalcante[1], **Snehal Patil**[1], **Yongseok Park**[2], **Laura S. Rozek**[3], and **Maureen A. Sartor**[1]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

[2]Department of Biostatistics, University of Pittsburgh, Pittsburgh, USA

[3]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, MI

Abstract

DNA methylation (5mC) plays important roles in mammalian development, oncogenesis, treatment response, and responses to the environment. DNA hydroxymethylation (5hmC) is also an informative epigenetic mark with distinct roles in regulation and cancer. Gold-standard, widely used technologies (bisulfite-conversion followed by deep sequencing) cannot distinguish between 5mC and 5hmC. Therefore, additional experiments are required to differentiate the two marks, and *in silico* methods are needed to analyze, integrate, and interpret these data. We developed the Methylation INTegration (mint) pipeline to support the comprehensive analysis of bisulfite-conversion (BS) and immunoprecipitation (IP) based methylation and hydroxymethylation assays, with additional steps towards integration, visualization and interpretation. The pipeline is available as both a command line and a Galaxy graphical user interface (GUI) tool. Both implementations require minimal configuration while remaining flexible to experiment specific needs.

Keywords

epigenomics; analysis pipeline; bisulfite sequencing; annotation; visualizations

Introduction

Methylation of cytosines to form 5-methylcytosine (5mC), especially at CpGs, is an epigenetic mark with important roles in mammalian development and tissue specificity, genomic imprinting, and environmental responses (1). Dysregulation of 5mC causes aberrant gene expression, affecting cancer risk, progression and treatment response (2). 5-hydroxymethylcytosine (5hmC) is an intermediate in the cell's active DNA demethylation pathway with tissue-specific distribution affecting gene expression (3) and carcinogenesis (4).

Corresponding Author: Maureen A. Sartor, University of Michigan, 100 Washtenaw Ave., 2017 Palmer Commons, Ann Arbor, MI 48109, 734-763-8013, sartorma@umich.edu.

Conflict of Interest Statement: The authors declare no potential conflicts of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Bisulfite-conversion (BS) assays such as whole genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) are widely used to quantify methylation levels at CpG-resolution. However, neither distinguishes 5mC from 5hmC since they are both protected from transformation under sodium bisulfite treatment. To distinguish the marks, OxBS-seq and TAB-seq detect either 5mC or 5hmC at CpG-resolution, respectively, however neither has been widely adopted. Immunoprecipitation (IP) assays such as MeDIP-seq, hMeDIP-seq, and hMeSeal are region-resolution assays detecting 5mC or 5hmC, respectively, and are more widely used and easily adopted. A review of these technologies can be found in (5).

Currently, differentiating 5mC from 5hmC is done *in silico* via signal integration from multiple assays. A small number of methylation analysis pipelines exist, though none support the integrative analysis of diverse data types for genome-wide 5mC and 5hmC. The methylPipe Bioconductor package (6) supports the analysis and visualization of base pair resolution methylation data. However, methylPipe does not support IP approaches and starts with bismark alignments, meaning users must perform QC, trimming, and alignment separately. SMAP is another methylation analysis pipeline, which supports processing from raw data, but only for RRBS experiments (7). Here we present mint, a methylation integration pipeline for processing, analyzing, integrating, and visualizing genome-wide 5mC and/or 5hmC data.

Overview of the mint pipeline

We developed mint to help jointly analyze and integrate 5mC and 5hmC genome-wide. From raw reads to integration and interpretation, users can analyze BS and IP technologies together (a ‘hybrid’ experiment), multiple IP technologies (e.g., MeDIP-seq and hMeDIP-seq ‘pulldown’ experiments), or a single type of experiment without integration.

For any experimental setup, mint performs quality control, adapter and quality trimming, sample-wise methylation quantification, and differential methylation analysis steps (Figure 1A-C). 5mC and 5hmC data are then integrated in a genomic segmentation based on overlapping signal, and genomic annotations with graphical summaries (Figure 1D-E) and a UCSC Genome Browser track hub (Figure 1F) are generated for seamless visualization, interpretation, and hypothesis-generation. The mint pipeline is implemented as a command-line tool using *make* (https://github.com/sartorlab/mint) (See Video 1 for introduction and installation), and as a GUI tool using the Galaxy web-based platform (8) (https://github.com/sartorlab/mint_galaxy).

Implementation and usage

Analysis with mint is modular (described below), with the 5mC (either BS or IP-based) and 5hmC (IP-based) data handled independently until the integration module. To setup a project in the command-line tool, users must create tables containing sample metadata (Table S1), and comparison metadata (Table S2) (See Video 2). After initializing a project, users may alter tool parameters in the *make* configuration file. In Galaxy, users input metadata and select appropriate files within the GUI. Tool parameters are specified on each tool’s Galaxy

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

page (Figure S1), and the tools are arranged into workflows. Galaxy workflows function both as pipelines and visualizations for the modules (Figure S2). The Galaxy implementation is currently limited to group versus group comparisons, with a planned future update to allow other experimental designs.

We demonstrate the mint pipeline on enhanced RRBS and hMeSeal data from two Acute Myeloid Leukemia (AML) samples with IDH2 mutations and two normal bone marrow (NBM) samples from (4) (GEO accession GSE52945). Previous findings indicate mutations in IDH2 lead to increased 5mC levels and decreased 5hmC levels, caused by an inhibition of the active demethylation process. In total, this data set has 8 pulldown samples and 4 bisulfite samples, and requires about 12 hours to run from raw reads to integration and visualization using 20 cores. Runtimes for other data will vary depending on the number samples, the number of CpGs covered, and available computing resources.

Alignment modules (pulldown_align and bisulfite_align

The alignment modules assess sample quality with FastQC, perform adapter and quality trimming with Trim Galore!, and align reads with bismark (9) for BS data and bowtie2 (10) for IP data. The reports for each sample are collated with MultiQC (11).

Sample modules (pulldown_sample and bisulfite_sample

The sample modules determine CpG-specific percent methylation levels for BS data with bismark methylation extractor (9) and qualitative methylation for IP data in the form of peaks called by macs2 (12). For each data type, mint performs ‘simple classifications’ of methylation levels into no, low, medium, or high. For BS data, thresholds of the absolute methylation level are used; for IP data, sample-wise tertiles based on fold change are used.

In our AML samples, we saw more hydroxymethylation peaks in IDH2 mutants, as previously reported (4). We observe more hydroxymethylation and less methylation in enhancers and 5’UTRs, respectively, compared to background regions (Figures S3A and S3B). We also observe hydroxymethylation to be similarly distributed across CpG features regardless of strength (Figure S3C), while we observe an increasing proportion of CpG island regions and decreasing CpG shelves as methylation strength decreases (Figure S3D).

Comparison modules (pulldown_compare and bisulfite_compare

The comparison modules test for differentially methylated CpGs (DMCs) or regions (DMRs) and differentially hydroxymethylated regions (DhMRs) with multi-factor designs allowing for categorical and/or continuous covariates (Table S2). For BS data, we allow users to destrand and group CpGs into tiles prior to testing for DMCs or DMRs with the R Bioconductor package DSS (13). The user sets FDR and methylation difference thresholds in the configuration file (or the Galaxy tool page) as criteria for differential methylation. For IP data, the R Bioconductor package csaw (14) tests for DhMRs, and the results are classified into weak, moderate, or strong DhMRs.

As in previous findings, we observe that mutations in IDH2 increase 5mC levels genome-wide (Figure S4A) and cause hypo-hydroxymethylation at specific loci, including the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

KIRREL locus (Figure S4B) (4). Additionally, hypo-hydroxymethylated regions in IDH2 samples tend to occur more often at 5' ends of genes and in exons with concurrent hyper-methylated regions at the same genomic annotations (Figures S4C and S4D). Interestingly, enhancers appear to be enriched for regions of hyper-hydroxymethylation and hyper-methylation in IDH2 samples (ibid). A pattern similar to the 5' ends of genes and exons is observed for regions at or near CpG islands (Figure S4E).

Integration modules (sample_classification and compare_classification

The integration modules segment the genome by 5mC and 5hmC signal per sample on the basis of overlapping signal, and/or by differential 5mC and 5hmC signal per comparison (Tables S3). For example, as in the sample module, integration of 5mC and 5hmC in the IDH2mut_2 sample shows that low levels of 5mC occur in very different regions relative to CpG islands than either 5hmC or high 5mC (Figure S4F). Integrating the DMRs and DhMRs from the IDH2 mutant versus NBM comparison reveals regions of joint differential methylation and hydroxymethylation, and we observe that regions of hyper-5mC and hypo-5hmC (with respect to IDH2 mutation) occur primarily at CpG islands, shores, and shelves (Figure S4G), as well as in promoters and exons of genic regions (Figure S4H).

Annotation and Genome Browser Tracks

To facilitate hypothesis generation and biological interpretation, results from each module are annotated to genomic features using the annotatr Bioconductor package (15). Default genomic features include CpG features (islands, shores, shelves, and open sea), genic features (1–5kb upstream of TSS, promoter (<1kb upstream of TSS), 5'UTR, exons, introns, and 3'UTR), enhancers, and lncRNAs.

The R session, summary tables, and plots tailored to the input data are saved, and users may reload them to further investigate the genomic annotations, summarize the data differently, alter default plots, or generate new plots (Figures 1D-E, and Figures S4 – S6). UCSC Genome Browser tracks are generated and arranged in a track hub folder for seamless viewing (Figures 1F and S5) (See Video 3 for an overview of results, annotations, and browser tracks).

Discussion

We developed the mint pipeline to jointly analyze 5-methylcytosine and 5-hydroxymethylcytosine signals *in silico* to better understand the biological roles of each epigenetic mark. The pipeline enables users to focus on optimizing parameters and interpreting experiments rather than interfacing with ten or more tools. The genomic annotations and default graphical outputs enable users to discover enriched features and associations that may have otherwise gone unexplored (e.g. overall hypo-methylation of CpG islands intersecting introns). Thus, mint streamlines data exploration and hypothesis generation, leading to discoveries that might otherwise be overlooked.

The modular design of mint facilitates 5mC and 5hmC integration, but also supports analysis of WGBS, RRBS, hMe-DIP, or Me-DIP, etc. experiments alone. Users can run mint on small pilot data and later add more samples without having to rerun previous samples.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Furthermore, its modularity allows users to stop and extract data at any step, and continue with a different program. If oxBS-seq and TAB-seq become more widely adopted, implementing support for them will be straightforward due to mint's modular implementation.

Here we presented mint using a small AML dataset, but are also using mint to analyze 5mC and 5hmC in a set of 36 head and neck squamous cell carcinoma samples, and experiments studying bisphenol-A effects on aging in mice. Regardless of context, the mint pipeline facilitates complex, comprehensive analyses of genome-wide methylation and hydroxymethylation data, enabling new biological discoveries.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Maria Figueroa for her helpful suggestions in the development of the mint pipeline.

Grant support: This work was funded by National Cancer Institute grant number R01 CA158286-S1 (S. Patil, L.S. Rozek, and M.A. Sartor) and the National Institute of General Medical Sciences Bioinformatics Training Grant (T32 GM070499) (R.G. Cavalcante).

Abbreviations

5mC	DNA methylation
5hmC	DNA hydroxymethylation
BS	Bisulfite-conversion
IP	Immunoprecipitation
WGBS	Whole genome bisulfite sequencing
RRBS	Reduced representation bisulfite sequencing
AML	Acute myeloid leukemia
NBM	Normal bone marrow

References

1. Schübeler D. Function and information content of DNA methylation. Nature. 2015; 517:321–6. [PubMed: 25592537]
2. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer. 2011; 11:726–34. [PubMed: 21941284]
3. Branco M, Ficz G, Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nat Rev Gen. 2012; 13:7–13.
4. Rampal R, Alkalin A, Madzo J, Vasanthakumar A, Pronier E, Patel J, et al. DNA Hydroxymethylation Profiling Reveals that WT1 Mutations Result in Loss of TET2 Function in Acute Myeloid Leukemia. Cell Reports. 2014; 9:1841–56. [PubMed: 25482556]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Song C-X, Yi C, He C. Mapping recently identified nucleotide variants in the genome and transcriptome. Nat Biotechnol. 2012; 30:1107–16. [PubMed: 23138310]
6. Kishore K, de Pretis S, Lister R, Morelli MJ, Bianchi V, Amati B, et al. methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. BMC Bioinformatics. 2015; 16:313. [PubMed: 26415965]
7. Gao S, Zou D, Mao L, Zhou Q, Jia W, Huang Y, et al. SMAP: a streamlined methylation analysis pipeline for bisulfite sequencing. GigaScience. 2015; 4:29. [PubMed: 26140213]
8. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016; 44(W1):W3–W10. [PubMed: 27137889]
9. Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011; 27:1571–2. [PubMed: 21493656]
10. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–9. [PubMed: 22388286]
11. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016; 32:3047–8. [PubMed: 27312411]
12. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012; 7:1728–40. [PubMed: 22936215]
13. Park Y, Wu H. Genome analysis Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016; 32:1446–53. [PubMed: 26819470]
14. Lun ATL, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. 2016; 44(5):e45. [PubMed: 26578583]
15. Cavalcante RG, Sartor MA. annotatr: genomic regions in context. Bioinformatics. 2017; doi: 10.1093/bioinformatics/btx183

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

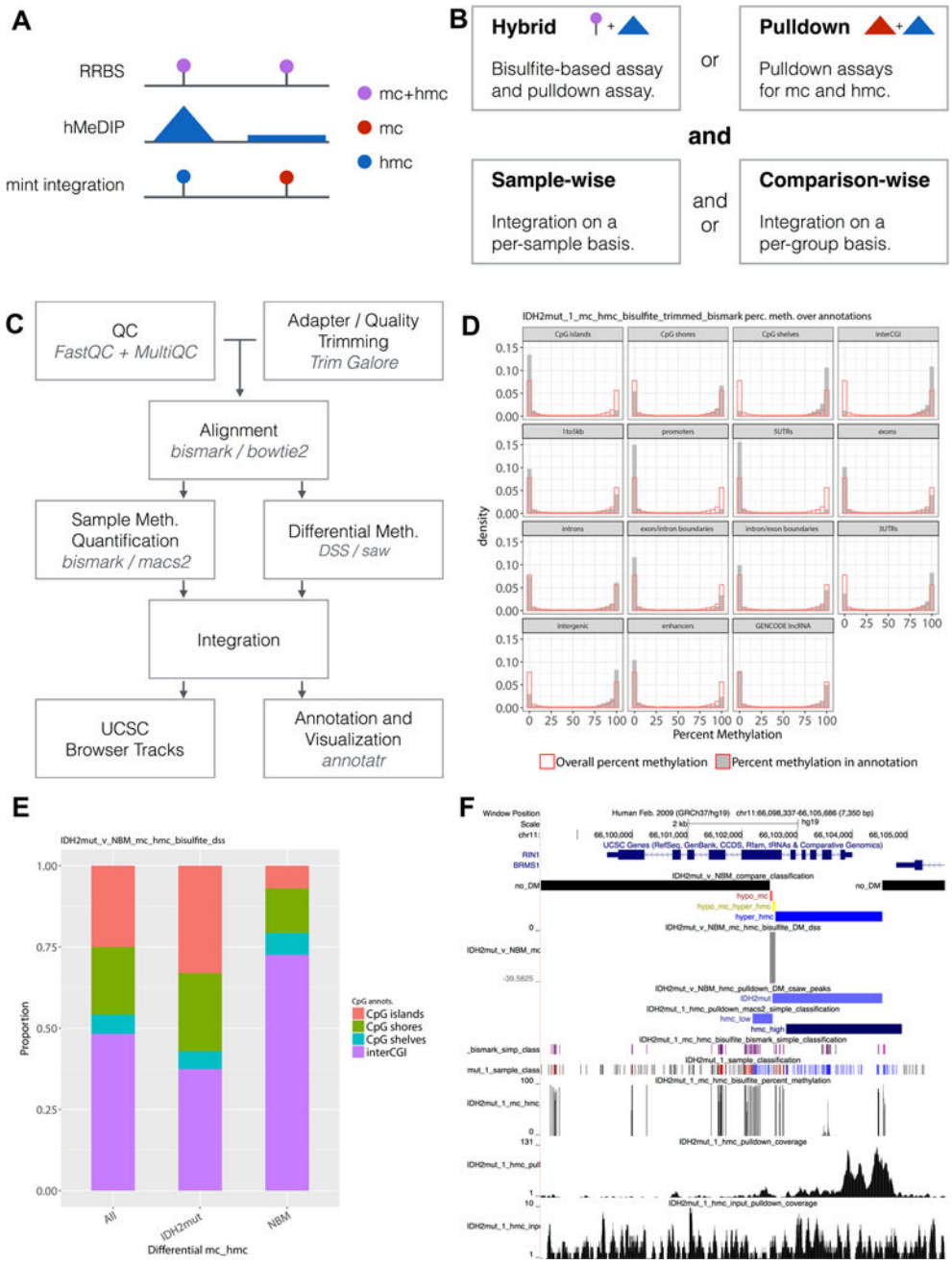

Figure 1. Overview of the mint pipeline, with selected outputs

(A) The concept behind 5mC and 5hmC integration. (B) A summary of supported experimental setups and analyses. (C) The overall mint workflow, with the primary tools used in each module. (D) and (E) are selected graphical outputs of genomic annotations. (D) The percent methylation in annotations from sample RRBS data. (E) The distribution of DhMRs from hMeSeal in CpG features. (F) Example UCSC Genome Browser tracks at the RIN1 locus showing coordinated hypo 5mC and hyper 5hmC (yellow) in an internal exon and hyper 5hmC at the 5' end (blue).