Examining and correcting low-mappability methylation calls in data from reads aligned with Bismark or Bwameth

Caiden Kumar, Brad Langhorst Ph.D.

August 2, 2019

Abstract

Methylation aligners, such as Bismark and Bwameth, frequently assign MapQ values to reads which are significantly higher than can be supported by the uniqueness of the region. These incorrectly high MapQs result in methylation calling in regions that are not sufficiently unique to support such calling. The result of this is that reads which should map to separate locations (possibly having different methylation states) actually end up mapping to the same locus, causing apparent mixed methylation at these loci. This can be fixed by filtering out insufficiently unique reads using Bismap mappability data. Simply filtering out Cs in insufficiently unique regions is not sufficient as it is prone to overfiltering in small mappability dips. Read filtering purifies much of the apparent mixed methylation to either 1 or 0. Specifically, examining Cs near genes containing ClinVar variants in a 50ng EM-seq data set from the NA12878 cell line reveals that 1405 Cs with apparent mixed methylation were purified to 0% methylation, and 2577 Cs with apparent mixed methylation were purified to 100% methylation.

Introduction

As DNA methylation status can have a significant biological function [21], it is important that there be an accurate way of calling methylation on a genome. Although there are multiple kinds of DNA methylation, a significant type is methylation of cytosine to 5-methylcytosine [1]. Data on DNA cytosine methylation state can be gathered using a methylation sequencing technique (see Figure 1), for example bisulfite sequencing [2]. In bisulfite sequencing, unmethylated cytosines are deaminated to uracil by the addition of sodium bisulfite. 5-methylcytosine is not affected. Since uracil sequences as thymine and 5-methylcytosine sequences as cytosine, positions of unmethylated Cs in a reference sequence can be identified by C->T transitions.

It is also possible to use an enzymatic method including TET2 to oxidize 5-methylcytosine and an APOBEC enzyme to deaminate unmodified cytosines

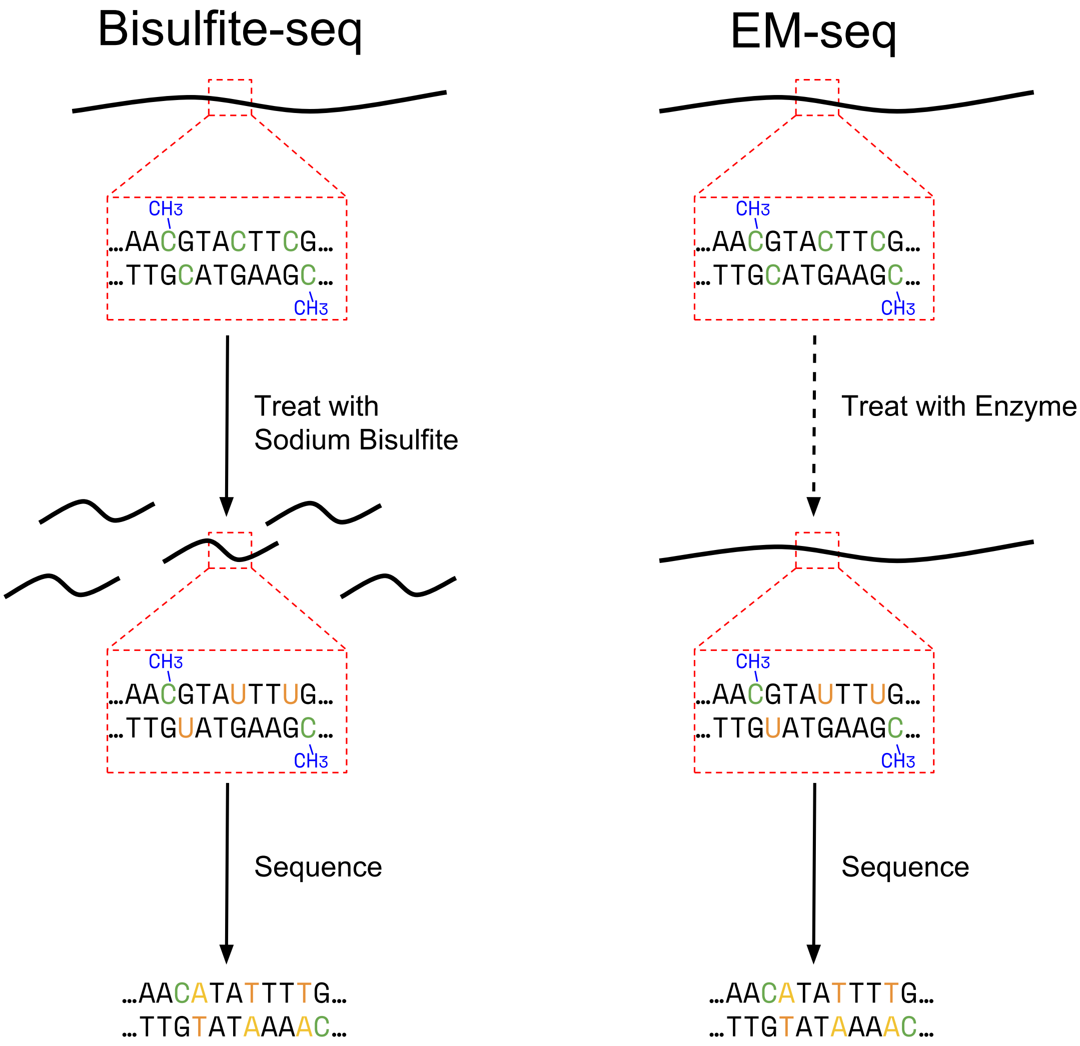

Figure 1: Overview of Methylation Sequencing Methods. <todo: Treat with Enzyme -> Tet2-> APOBEC

to uracil. While sodium bisulfite treatment produces other DNA damage, the enzymatic method deaminates with more precision (forthcoming publication).

Whichever method is used to deaminate cytosines, sequence data is typically aligned to a reference genome using a methylation-aware aligner [4], which is specifically designed to handle the C->T transitions in methylation sequencing data during alignment of the reads to a reference. Once aligned, the data can be passed through a methylation caller such as MethylDackel [20] or bismark_methylation_extractor [10], which will use the resulting reads to determine the methylation status of a particular cytosine (see Figure 2). The resulting data shows the methylation status of each cytosine in the genome and can therefore be used to find and study biologically significant DNA methylation sites.

A read must be unambiguously placed if it is to provide information about a specific locus. Reads that equally match more than one area of the reference genome should not be used to assess methylation of any given C. To avoid calling Cs using reads derived from multiple genome loci, methylation aligners (and read aligners in general) assign a MapQ value to each read alignment (see Figure 2). According to the SAM specification [5], MapQ is defined as: "$-10 \log_{10} \Pr\{\text{mapping position is wrong}\}$, rounded to the nearest integer". MapQ indicates how uniquely placed a read alignment is, that is, in how many other places could the read align to the reference. A low value of MapQ means that the read aligns in many places throughout the genome (for instance, a read of centromeric satellite DNA would likely have a very low MapQ). A high value of MapQ indicates that the read likely aligns where it is placed and nowhere else in the genome. For example, a MapQ of 10 means that that read would only have a 1:10 probability of being incorrectly placed.

However, Bismark (as a result of using Bowtie2 [13] for alignment) [11] reports MapQ based on number of reference mismatches instead, producing values between 0 and 42 <citation>. Bwameth (using bwa [14] for alignments) [18] follows the SAM specification in estimating probabilities in the MapQ field.

A methylation caller can use accurate MapQ values to filter out reads with multiple placements in the genome, allowing the resulting methylation calls to accurately reflect their specific loci.

Results

While evaluating two commonly used methylation aligners, Bismark [10] and Bwameth [17], we observed a significant number of reads with high MapQ values in repetitive regions (e.g. centromeres). While both aligners are affected, Bismark produced more high MapQ reads in repetitive regions (see Figure 3). After observing high MapQ reads in the centromere and other repetitive regions, we investigated to see if smaller regions might also be too repetitive to support the high aligner MapQ estimates observed. We identified repetitive regions using data from data from Bismap [8], a tool that counts the number of occurrences of every single K-mer of a particular length (in this case, k=100) in the genome

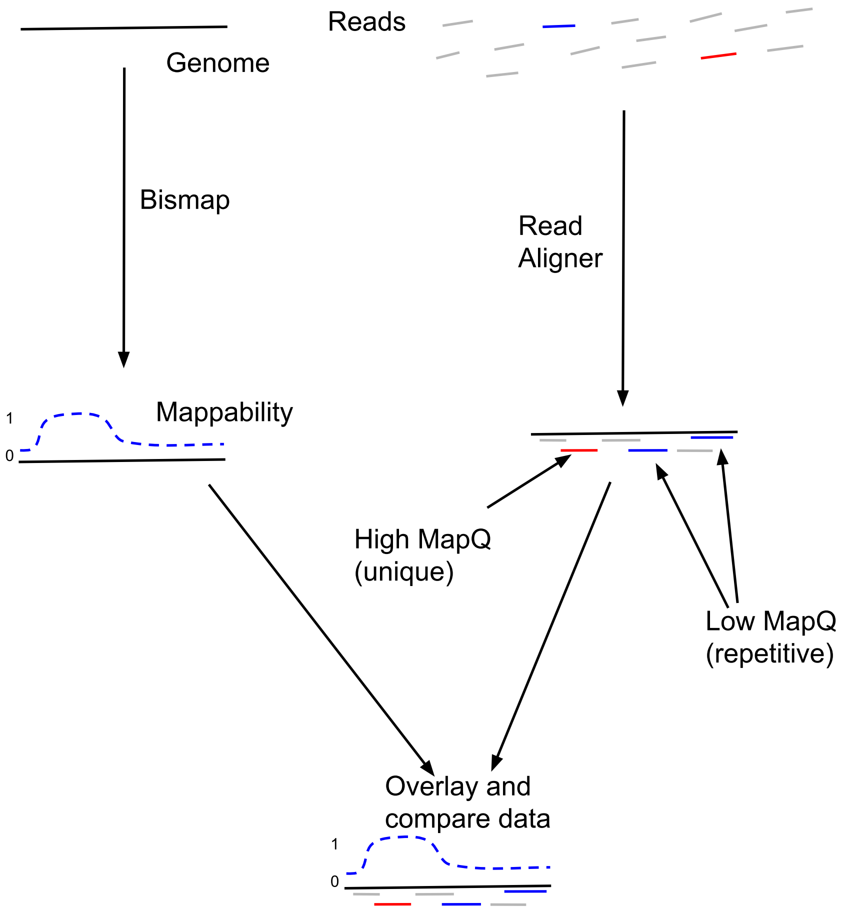

Figure 2: Experimental Overview

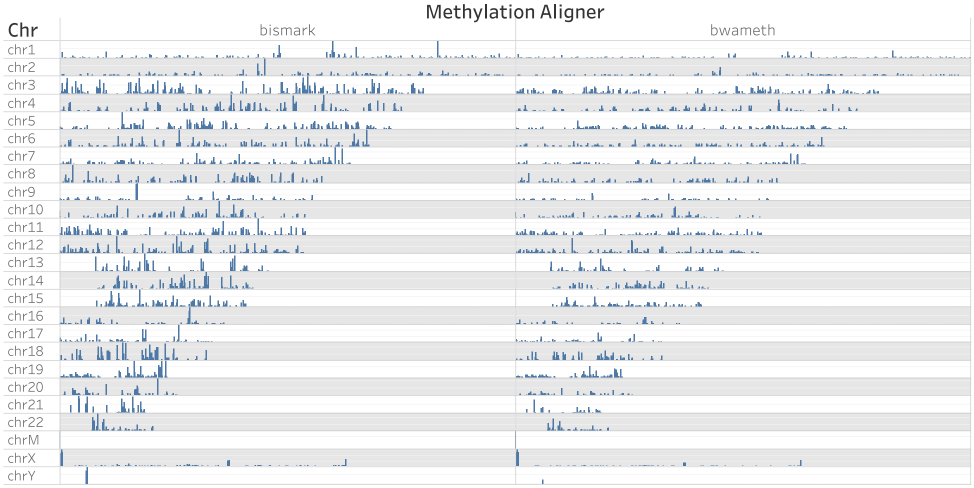

Figure 3: Example of MapQ Issue. On the Cov Difference track, red means that Bismark has higher coverage and blue means Bwameth has higher coverage.

to create a mappability score for every base in the Grch38 reference. <bases this on assumption that each locus is either 0 or 100%> Bismap takes the effect of C->T conversion into account. Reads entirely contained within a region of low mappability should not have high MapQ values, however we observed many high-MapQ reads in regions with very low or zero Bismap mappability (see Figure 4).

Bismark and Bwameth use different systems to define values for MapQ (see supplemental detail), but even stringent MapQ thresholds cannot reliably select reads for safe methylation calling in regions containing repetitive DNA. We considered excluding methylation calls on Cs found in low-mappability regions (e.g. using bedtools), but rejected this approach because it is prone to both over and under filtering. Cs in short, unique regions would be kept (underfiltered) even if the surrounding DNA is repetitive (see Figure 5). Overfiltering of Cs in short repetitive regions is a larger problem. In this scenario, a C located in a small dip in mappability would be eliminated (overfiltered) despite coverage from read pairs anchored in nearby unique regions (see Figure 7). In practice, underfiltering is rare (x% of Cs) but overfiltering is more common (y% of Cs) (see Figure 6). <coverage filtering can help if many loci are collapsed to few in reference, but if data is simply mixed between loci, it does not help>

To reliably filter only problematic read pairs (those where both mates are placed in low mappability regions) we modified the MethylDackel methylation caller to accept a bigWig file of low-mappability regions to exclude from analysis. This alignmentfiltering approach precisely eliminates only those reads in repetitive regions and does not incur a significant cost in terms of execution speed (? min with the patch ? min without).(see computational methods for

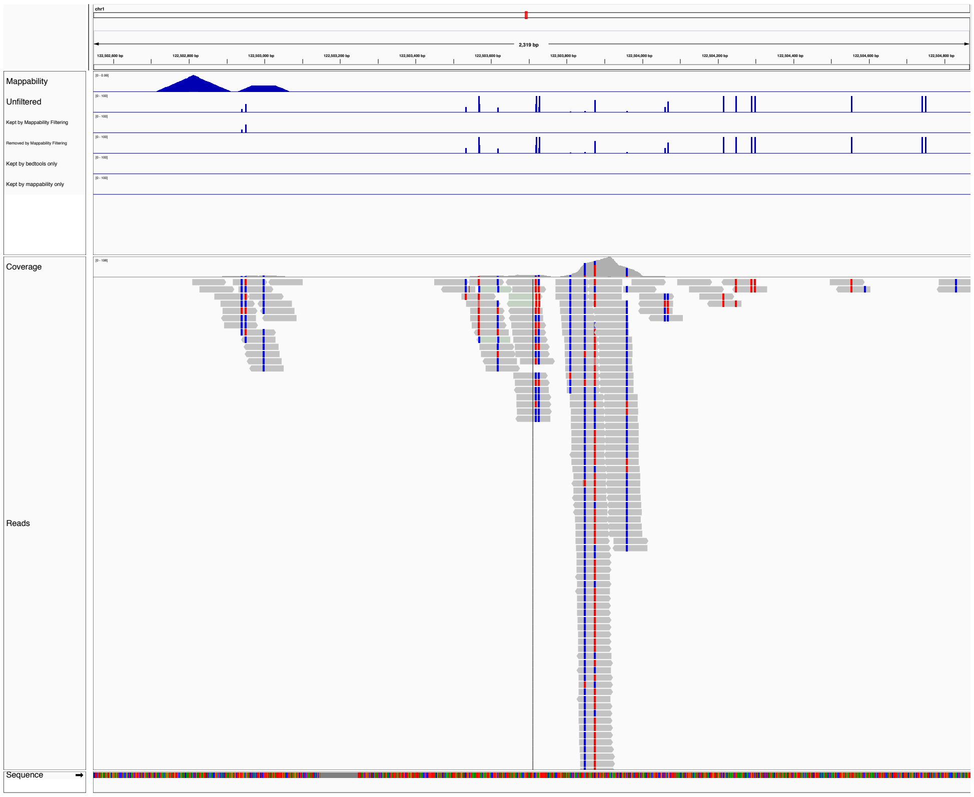

Figure 4: An example of how MapQ filtering does not remove all poorly mapped reads.

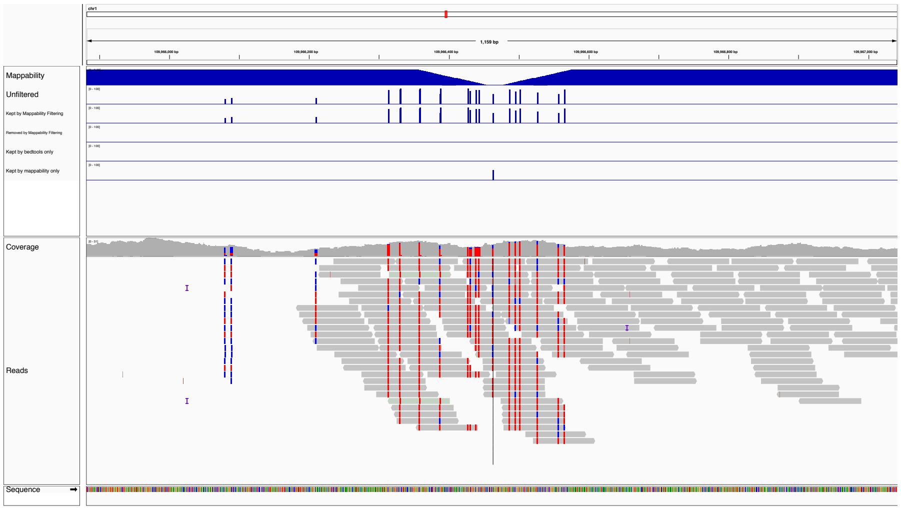

Figure 5: An example of a C incorrectly discarded by coordinate filtering.

details)

To focus on the incorrect alignment and methylation calls with biological and medical significance, we examined methylation calls in GENCODE genes [6] that contain variants listed in the ClinVar [12] database of disease-associated variants. Using our alignment filtering approach we successfully avoided calling X Cs in these important regions having insufficient mappability to support methylation calling. Briefly, we intersect read alignments and MapQ data with Bismap low mappability regions counting Cs in regions where the MapQ is higher than should be possible given the mappability (see Figure 2).

todo: filtering effect em-seq vs wgbs (fig) for bwameth

todo: filtering effect em-seq-vs wgbs for bismark (this is the data I'm working on)

todo: # cs affected (worst) bismark vs filtered vs bwameth vs filtered (best) (fig?) (I can get this from the emseq/wgbs data, it's a simple bedtools operation)

todo: group Cs by magnititude of change in methylation post filtering (fig, heatmap?) (best done by comparing both %methyl and read count, maybe we should discuss how to layout this figure though)

todo: effect on ClinVar

todo: examples of specific effects in ClinVar

todo: examine GEO dataset (maybe, depends on time/progress)

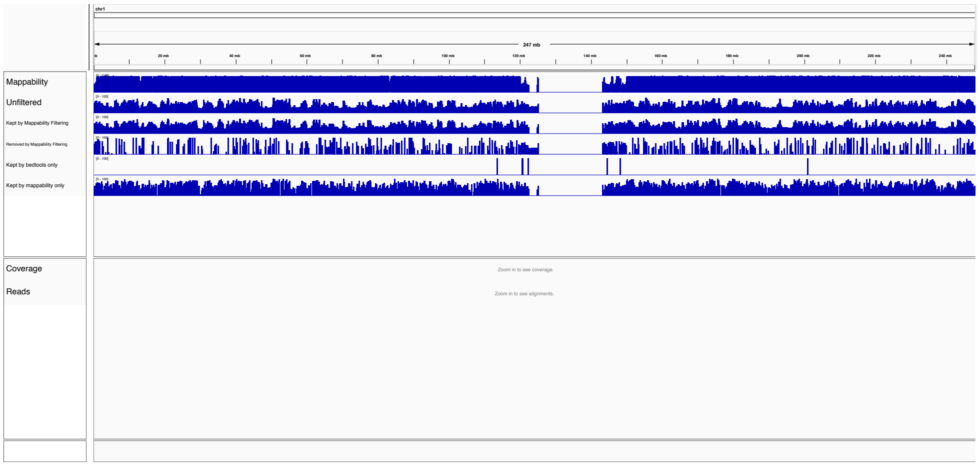

Figure 6: An overview of the occurrences of these scenarios in grch38 chromosome 1.

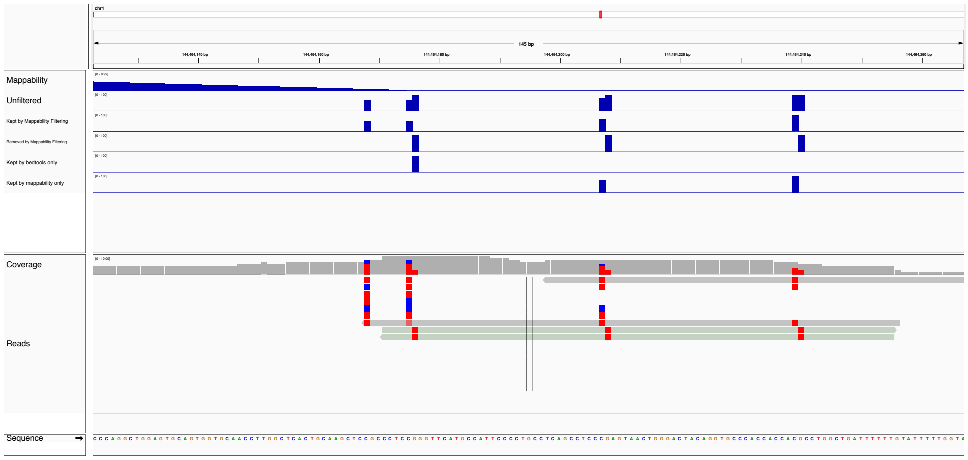

Figure 7: An example of a C incorrectly kept by coordinate filtering.

Materials and Methods

DNA Methylation Sequencing

Materials

Libraries were prepared from 200 ng of genomic DNA from the NA12878 cell line (Coriel). This input was supplemented with a small amount of fully-methylated Xp-12 DNA, lambda phage DNA (NEB #N3011) and a pUC19 plasmid (NEB #N3041) treated with M.SssI CpG Methyltransferase (NEB #M0226).

Whole Genome Bisulfite Libraries

Libraries were prepared using the Ultra II DNA library prep kit before being Bisulfite converted according to (protocol).

Sequencing

Libraries were pooled and sequenced with diverse libraries (~10%) on 2 flow cells of an Illumina Novaseq 6000 [7] using the S2 chemistry<citation>. We acquired 1.55 billion 100bp paired-end reads for the EM-seq method and 1.60 billion paired-end reads for the Bisulfite converted libraries.

Computational Methods

In order to compare the two aligners, this analysis was run twice: once with Bismark, once with Bwameth. In order to run the analysis, the following data and tools were used:

The GRCh38.p11 analysis set supplemented with phage T4, phage lambda, phage Xp12 and pUC19 contigs was used throughout [3]. A VCF file of disease-associated variant sites was downloaded from ClinVar (todo: check this clinvar_20190722.vcf.gz) and the GENCODE v31 gene annotations were used. The mappability data was the 100bp multi-track bigWig file downloaded from Bismap (see supplemental materials for link). A detailed diagram of the process is found in Figures 8 and 9. Tools used include bedtools [19], samtools [15], GNU awk, MethylDackel, bigWigToBedGraph [9], BEDOPS [16], GNU sort, GNU head, bwameth, and bismark. GNU sort and head are required since the functionality needed (the ability to parallelize sorting, and to use a negative value with the -n option for head to count lines from the end of the file) is not present in BSD sort and head.

Methylation calling was performed on the aligned BAMs using MethylDackel. Methyldackel uses a default value of 10 as a minimum MapQ which should include mostly single locus reads for both aligners, however inaccurate MapQs lead to reads being incorrectly included in methylation calling. In order to eliminate reads in low-mappability regions, a patch was created for MethylDackel which allows it to take as an input a bigWig file which is then used to filter out read pairs (this patch currently only supports paired-end reads) where

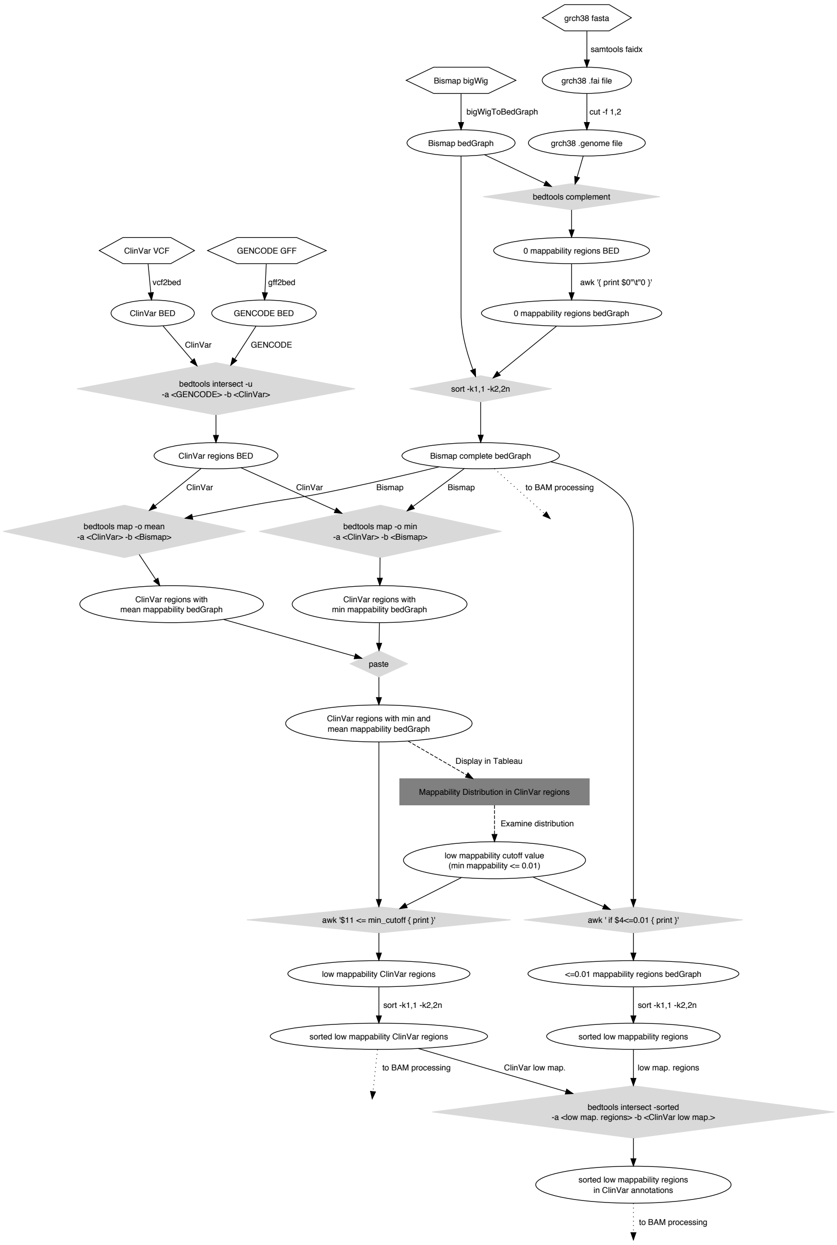

Figure 8: Detailed description of preprocesing steps in data analysis. <todo: add legend with figure features, boxes, ovals, edge types, etc. parts A and B with >

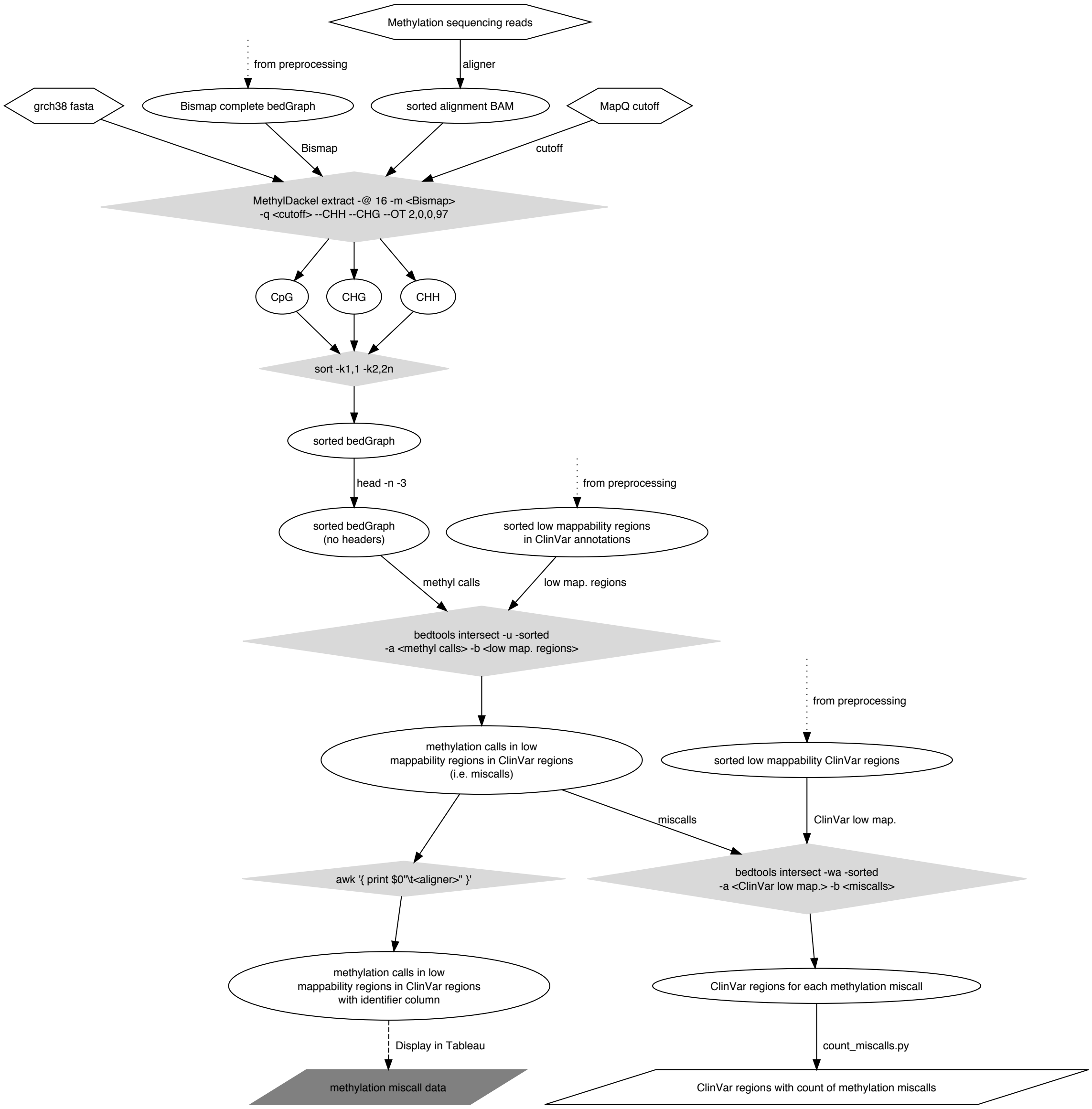
Methylation sequencing reads
from preprocessing
aligner
grch38 fasta
Bismap complete bedGraph
sorted alignment BAM
MapQ cutoff
Bismap
cutoff
MethylDackel extract -@ 16 -m <Bismap>
-q <cutoff> --CHH --CHG --OT 2,0,0,97
CpG
CHG
CHH
sort -k1,1 -k2,2n
sorted bedGraph
head -n -3
from preprocessing
sorted bedGraph
(no headers)
sorted low mappability regions
in ClinVar annotations
methyl calls
low map. regions
bedtools intersect -u -sorted
-a <methyl calls> -b <low map. regions>
from preprocessing
methylation calls in low
mappability regions in ClinVar regions
(i.e. miscalls)
sorted low mappability ClinVar regions
miscalls
ClinVar low map.
awk '{ print $0"\t<aligner>" }'
bedtools intersect -wa -sorted
-a <ClinVar low map.> -b <miscalls>
methylation calls in low
mappability regions in ClinVar regions
with identifier column
ClinVar regions for each methylation miscall
Display in Tableau
count_miscalls.py
methylation miscall data
ClinVar regions with count of methylation miscalls

Figure 9: Detailed description of BAM processing steps in data analysis. <todo: add legend with figure features, boxes, ovals, edge types, etc. parts A and B with >

neither mate intersects a high-mappability region. The patch allows for user configuration of the low mappability threshold and the number of bases which must be equal to or above that threshold in order for the read pair to be kept (the defaults are a low mappability threshold of 0.01 and to require 15 bases that are greater than or equal to that threshold in a single read). The filtering algorithm has been optimized both by loading the mappability data into memory before calling, and through the use of a custom run-length compressed binary file format that we here term BBM (Binary BisMap). This format can store the mappbility data for the hg38/grch38 human genome in 143 MB, compared to 1.11 GB when stored as a bigWig, giving a compression ratio for this dataset of 7.78:1.

The ClinVar VCF and GENCODE GFF were combined using bedtools and BEDOPS into a single BED file listing all GENCODE annotation regions that overlap one or more ClinVar variants (these regions will be referred to as "ClinVar regions"), which will be called the "ClinVar regions BED". The Bismap bedGraph (which, as is standard for bedGraph files, did not contain zeroes) and the grch38 fa file were processed with bedtools, awk, and GNU sort to obtain a bedGraph containing all mappability data, including zeroes. This file will be referred to as the "Bismap complete bedGraph". It was then combined with the ClinVar regions BED using bedtools map to create a file containing the minimum and mean mappability for each gene with a ClinVar variant.

The Bismap complete bedGraph was then filtered using awk to produce a file containing only low mappability regions (mappability < 0.01). The file with minimum and mean mappability for every ClinVar region was filtered likewise on minimum mappability. The two resulting files (one of low-mappability regions, one of ClinVar regions with low minimum mappability) were combined to produce a file of all low mappability regions that are in ClinVar regions.

Alignments of 2x99-bp paired-end EM-seq and whole genome bisulfite reads were processed using MethylDackel (with and without the custom patch) using a minimum MapQ cutoff of 20 and the default settings mentioned above and combined with the file of all low mappability regions in ClinVar regions to produce a list of all miscalls (this will be referred to as the "miscalls file"). A miscall, as used here, is defined as a methylation call that was filtered out using the new MethylDackel mappability filtering.

The miscalls file was intersected with the file of ClinVar regions with low minimum mappability to produce a file of ClinVar regions with miscalls. The -wa option for bedtools intersect was used here, so this file contained duplicates. Specifically, it contained one copy of a ClinVar region for each miscall in that region. These duplicates were then used to count miscalls by feeding the data to a custom Python script (see the supplemental materials) which counted and combined the duplicates, producing a list of all ClinVar regions with miscalls and how many miscalls are in the region.

Since this analysis was run for both Bismark and Bwameth, the miscalls file was also processed through awk to add a field specifying which aligner the data is from. The same field was added to the list of all ClinVar regions with miscalls and counts described previously. Both files were examined and compared in

Tableau®.

<still getting data>

Discussion

Reads placed with falsely high confidence have cascading detrimental effects on methylation calling, differential methylation assessment, and assement of phenotypes associated with methylation status. Because of the more accurate MapQ values, decreased run time and more flexibility to separate methylation calling from alignment, we recommend the use of bwameth for alignment and MethylDackel with MapQ $>$ 20 for methylation calling. To further improve accuracy of methylation asssessment, reads with both mates in low mappability regions should be excluded.

<waiting on results>

Supplemental Materials

The Nextflow script used to analyze this data can be found at <to be added>. In order to run, it requires vcf2bed, gff2bed, bigWigToBedGraph, cut, MethylDackel (whatever-version or newer), GNU sort, GNU head, bedtools, paste, awk, and Python 3 <need to cite those of these tools which haven't been cited earlier>.

The pull request for the patch adding mappability support to MethylDackel can be found at `https://github.com/dpryan79/MethylDackel/pull/80`. It was merged into MethylDackel in version (whatever-version).

The Bismap file used for this analysis was downloaded from `https://www.pmgenomics.ca/hoffmanlab/proj/bismap/trackhub/hg38/k100.Bismap.MultiTrackMappability.bw`

References

[1] Achim Breiling and Frank Lyko. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & Chromatin*, 8:24, 2015.

[2] Gary G. Chen, Jeffrey A. Gross, Pierre-Eric Lutz, Kathryn Vaillancourt, Gilles Maussion, Alexandre Bramoulle, Jean-François Théroux, Elena S. Gardini, Ulrike Ehlert, Geneviève Bourret, Aurélie Masurel, Pierre Lepage, Naguib Mechawar, Gustavo Turecki, and Carl Ernst. Medium throughput bisulfite sequencing for accurate detection of 5-methylcytosine and 5-hydroxymethylcytosine. *BMC Genomics*, 18:96, Jan 2017. 28100169[pmid].

[3] Deanna M. Church, Valerie A. Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M. McLaren, Graham R. S. Ritchie, Derek Albracht, Milinn

Kremitzki, Susan Rock, Holland Kotkiewicz, Colin Kremitzki, Aye Wollam, Lee Trani, Lucinda Fulton, Robert Fulton, Lucy Matthews, Siobhan Whitehead, Will Chow, James Torrance, Matthew Dunn, Glenn Harden, Glen Threadgold, Jonathan Wood, Joanna Collins, Paul Heath, Guy Griffiths, Sarah Pelan, Darren Grafham, Evan E. Eichler, George Weinstock, Elaine R. Mardis, Richard K. Wilson, Kerstin Howe, Paul Flicek, and Tim Hubbard. Modernizing reference genome assemblies. *PLoS biology*, 9(7):e1001091, July 2011.

[4] Francine E. Garrett-Bakelman, Caroline K. Sheridan, Thadeous J. Kacmarczyk, Jennifer Ishii, Doron Betel, Alicia Alonso, Christopher E. Mason, Maria E. Figueroa, and Ari M. Melnick. Enhanced reduced representation bisulfite sequencing for assessment of dna methylation at base pair resolution. *J Vis Exp*, (96):52246, Feb 2015. 25742437[pmid].

[5] The SAM/BAM Format Specification Working Group. Sequence Alignment/Map Format Specification, May 2018. https://samtools.github.io/hts-specs/SAMv1.pdf.

[6] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. Gencode: The reference human genome annotation for the encode project. *Genome Res*, 22(9):1760–1774, Sep 2012. 22955987[pmid].

[7] Illumina. Novaseq 6000. https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html.

[8] Mehran Karimzadeh, Carl Ernst, Anshul Kundaje, and Michael M. Hoffman. Umap and bismap: quantifying genome and methylome mappability. *Nucleic Acids Research*, 2018.

[9] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. Bigwig and bigbed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, 2010. http://genome.ucsc.edu/.

[10] Felix Krueger and Simon R. Andrews. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.

[11] Felix Krueger and Simon R. Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, June 2011.

[12] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018.

[13] Ben Langmead, Christopher Wilks, Valentin Antonescu, and Rone Charles. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, 35(3):421–432, February 2019.

[14] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595, March 2010.

[15] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[16] Shane Neph, M. Scott Kuehn, Alex P. Reynolds, Eric Haugen, Robert E. Thurman, Audra K. Johnson, Eric Rynes, Matthew T. Maurano, Jeff Vierstra, Sean Thomas, Richard Sandstrom, Richard Humbert, and John A. Stamatoyannopoulos. Bedops: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920, 2012.

[17] B. S. Pedersen, K. Eyring, S. De, I. V. Yang, and D. A. Schwartz. Fast and accurate alignment of long bisulfite-seq reads. *ArXiv e-prints*, January 2014.

[18] Brent S. Pedersen, Kenneth Eyring, Subhajyoti De, Ivana V. Yang, and David A. Schwartz. Fast and accurate alignment of long bisulfite-seq reads. *arXiv:1401.1129 [q-bio]*, January 2014. arXiv: 1401.1129.

[19] Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[20] Devon Ryan. MethylDackel. https://github.com/dpryan79/MethylDackel.

[21] Dirk Schübeler. Function and information content of dna methylation. *Nature*, 517:321 EP –, Jan 2015.