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

Caiden Kumar, Brad Langhorst Ph.D.

August 15, 2018

Abstract

<abstract>

Background

As DNA methylation status can have a significant biological function [17], 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 2), 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 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 [16] or bismark_methylation_extractor [10], which will use the resulting reads to determine the methylation status of a particular cytosine (see Figure 1). 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 are 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 1). 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.

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 [14], we observed a significant number of reads with high MapQ values in repetitive regions (e.g. centromeres). While both aligners are affected, Bismark produced significantly 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 to create a mappability score for every base in the Grch38 reference. Bismap takes the effect of C->T transitions into account. Reads entirely contained within a region of low mappability should not have high MapQ values.

We observed many high-MapQ reads in regions with very low or zero Bismap mappability (figure). According to the SAM specification, MapQs of 20 should only have a 1:100 probability of being incorrectly placed, which is not supported by the region's Bismap mappability. Bismark reports MapQ based on number of reference mismatches, producing values between 0 and 42. Bwameth (using bwa alignments) follows the SAM specification in estimating probabilities in the MapQ field. 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 focus on the incorrect alignment and methylation calls with biological and medical significance, this analysis examines calls in GENCODE genes [6] that contain variants listed in the ClinVar [11] database of disease-associated variants. 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 1).

Problematic read pairs are those where both mates are placed in low mappability regions. To reject only problematic pairs, we modified the MethylDackel methylation caller to accept a bigWig file of low-mappability regions to exclude from analysis. We do not want to simply exclude reads overlapping a low mappability region because the longer insert sizes now possible with enzymatic methylation sequencing may allow placement of one mate in a unique region and allow assessment of more loci in repetitive regions. This simple filtering approach incurs a significant cost as we query the bigWig file for each read, but we did not want to delay reporting this observation while we improve the algorithm's performance.

todo: filtering effect em-seq vs wgbs

todo: # cs affected (worst) bismark vs filtered vs bwameth vs filtered (best)

todo: group Cs by magnitude of change in methylation post filtering

Materials and Methods

DNA Methylation Sequencing

Materials

Libraries were prepared from 200 ng of genomic DNA from the NA12878 cell line<source>. This input was supplemented with a small amount of fully-methylated Xp-12 DNA<source> (todo: confirm that both libs have XP-12), lambda phage DNA (NEB #N3011)<source>, 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.

Sequencing

Libraries were pooled and sequenced with diverse libraries (~10%) on 2 flow cells of an Illumina Novaseq 6000 [7] using the S2 chemistry<citatkino>. We acquired 1,554,674,639 100bp paired-end reads for the EM-seq method and 1,601,169,878 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:

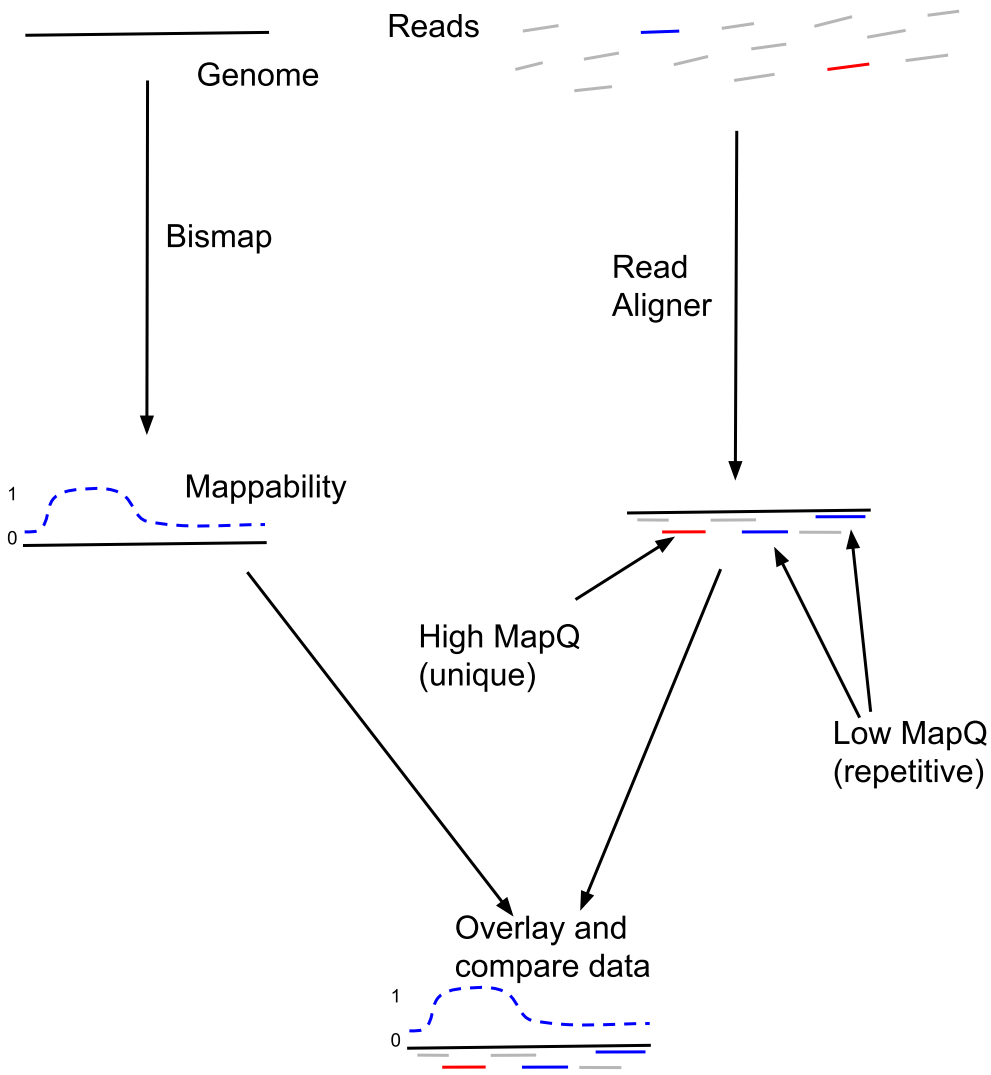

Figure 1: Experimental Overview

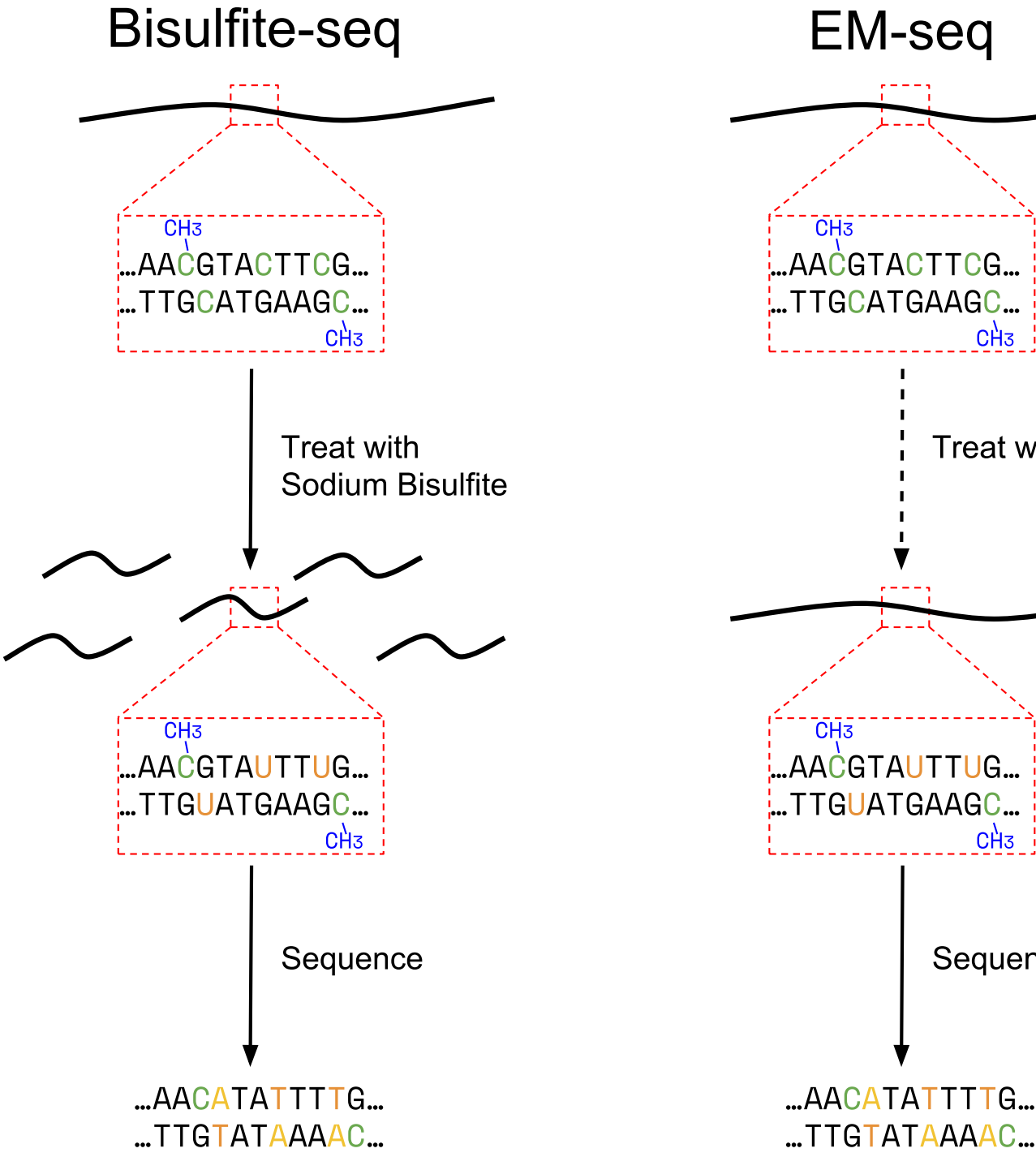

Figure 2: Overview of Methylation Sequencing Methods. <todo: Treat with Enzyme -> Tet2-> APOBEC>

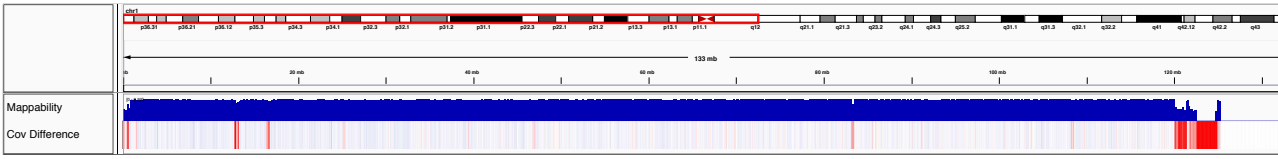

Figure 3: Example of MapQ Issue. On the Cov Difference track, red means that Bismark has higher coverage and blue means Bwameth has higher coverage.

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_20180729.vcf.gz) and the GENCODE v28 gene annotations were used. The mappability data was the 100bp multi-track bigWig file downloaded from Bismap (https://bismap.hoffmanlab.org/raw/hg38/k100.bismap.bedgraph.gz). A detailed diagram of the process is found in Figures 4 and 5. Tools used include bedtools [15], samtools [12], GNU awk, MethylDackel, bigWigToBedGraph [9], BEDOPS [13], GNU sort, GNU head, bwameth, and bismark. GNU sort and head are required since the functionality needed is not present in BSD sort and head.

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 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 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 above that threshold in a single read).

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 less than or equal to 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.

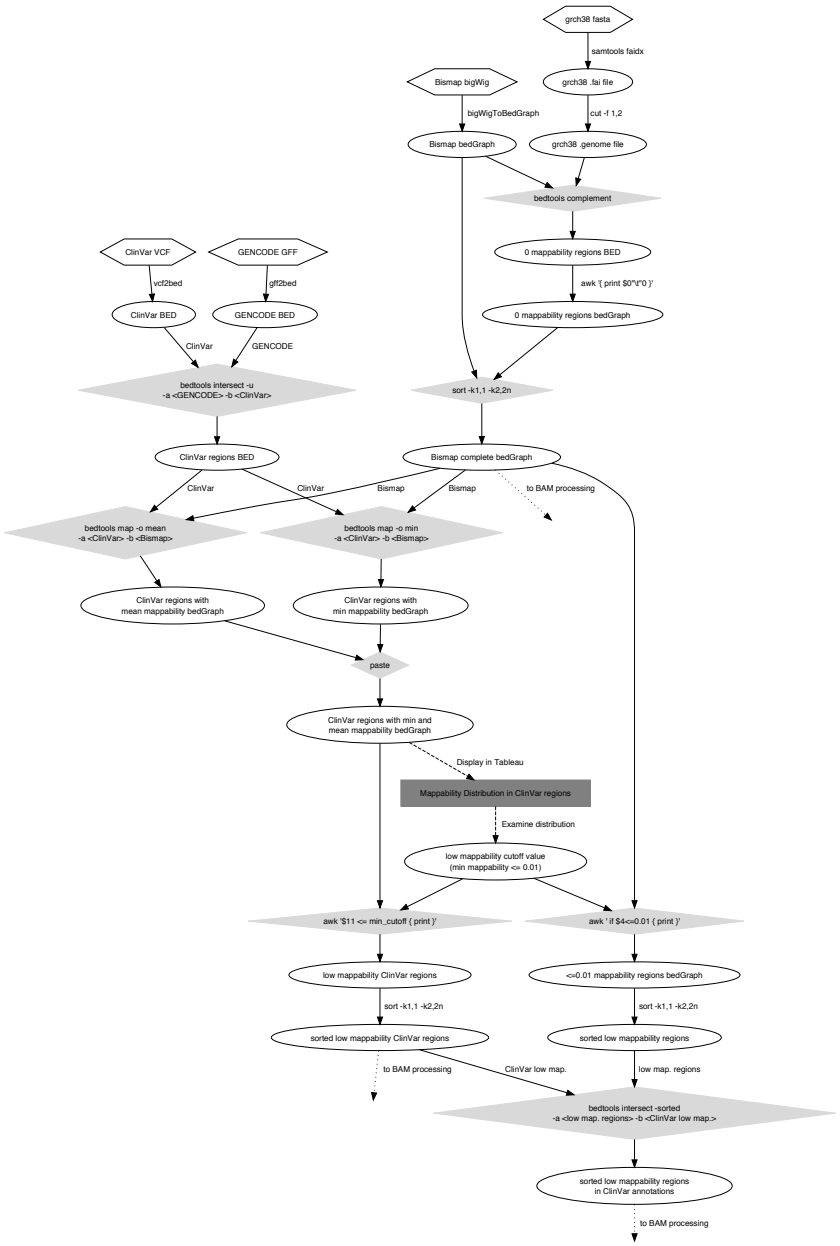

Figure 4: Detailed description of preprocesing steps in data analysis. <todo: add legend with figure features, boxes, ovals, edge types, etc. parts A and B with >

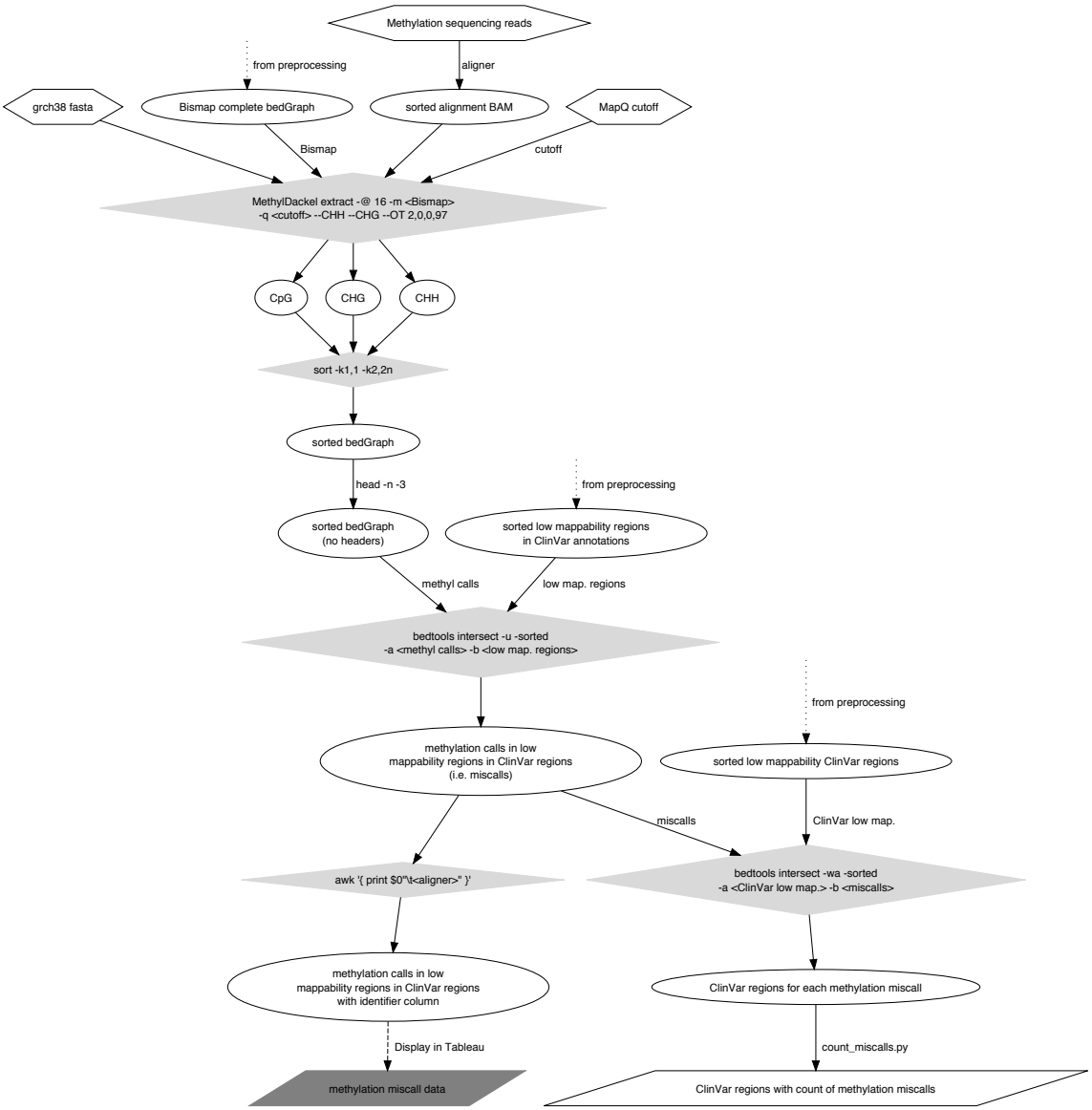

Figure 5: Detailed description of BAM processing steps in data analysis. <todo: add legend with figure features, boxes, ovals, edge types, etc. parts A and B with >

Alignments of 2x100-bp paired-end EM-seq 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 in a region which has Bismap mappability less than or equal to 0.01.

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®.

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 a low mappability regions should be excluded.

<waiting on results>

Supplemental Materials

<nextflow script>

Patch adding mappability support to MethylDackel

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] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

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