Correcting Methylation Calls in Low-Mappability Regions

Caiden M. Kumar, Ariel Erijman, Bradley W. Langhorst

June 30, 2021

Abstract

DNA methylation is an important component in vital biological functions such as embryonic development, carcinogenesis, and heritable regulation. Accurate methods to assess genomic methylation status are crucial to its effective use in many scenarios, especially in the detection and diagnosis of disease. Methylation aligners, such as Bismark and bwa-meth, frequently assign MapQ values to reads which are significantly higher than can be supported by the uniqueness of the region they are mapped to. These incorrectly high MapQs result in inappropriate methylation calling in repetitive regions. We observe reads that should map to separate locations (possibly having different methylation states) actually end up mapping to the same locus, causing apparent mixed methylation at such loci. Methylation calling can be improved by using Bismap mappability data to filter out insufficiently unique reads. Simply filtering out Cs in insufficiently unique regions is not adequate as it is prone to overfiltering Cs in small mappability dips. These Cs can in fact often be called using reads anchored in a nearby mappable region. We have created a patch for the MethylDackel methylation caller to perform read-based filtering. Read-based filtering resolves some of the apparent mixed methylation to either 0% or 100% methylation. We examined methylation calls with and without read-based filtering in or near the 7830 genes containing ClinVar variants in a methylation sequencing data set from the NA12878 cell line and in tumor samples. Examining low mappability Cs in the NA12878 data set revealed 1405 mixed methylation Cs were corrected to 0% methylation, and 2577 mixed methylation Cs were corrected to 100% methylation.

Introduction

As DNA methylation status can have a significant biological function [19], it is important that there be an accurate way of calling methylation on a genome. Although there are multiple varieties of DNA methylation, a significant type is methylation of cytosine to 5-methylcytosine [2]. Data on DNA cytosine methylation state can be gathered using a methylation sequencing technique

(see Figure 1), for example bisulfite sequencing [3]. In bisulfite sequencing, unmethylated cytosines are deaminated to uracil by the addition of sodium bisulfite. 5-methylcytosines are 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[3].

It is also possible to use an enzymatic method employing TET2 to oxidize 5-methylcytosine and an APOBEC enzyme to deaminate unmodified cytosines to uracil. While sodium bisulfite treatment produces other DNA damage, this 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 [5], which is specifically designed to handle the C->T transitions in methylation sequencing data when aligning the reads to a reference. Once aligned, the data can be passed through a methylation caller such as MethylDackel [18] or bismark_methylation_extractor [12], 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 genomic loci, methylation aligners (and read aligners in general) assign a MapQ value to each read alignment (see Figure 2). According to the SAM specification [6], 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 MapQ means that the read may align in many places throughout the genome (for instance, a read of centromeric satellite DNA would likely have a very low MapQ). A high 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 the methylation aligner bwa-meth [16], we observed a significant number of reads with unexpectedly high MapQ values in repetitive regions (e.g. centromeres). After observing these high MapQ reads in the centromere and larger 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 Bismap [10], a tool that counts the number of occurrences of every single K-mer of a particular length (in this

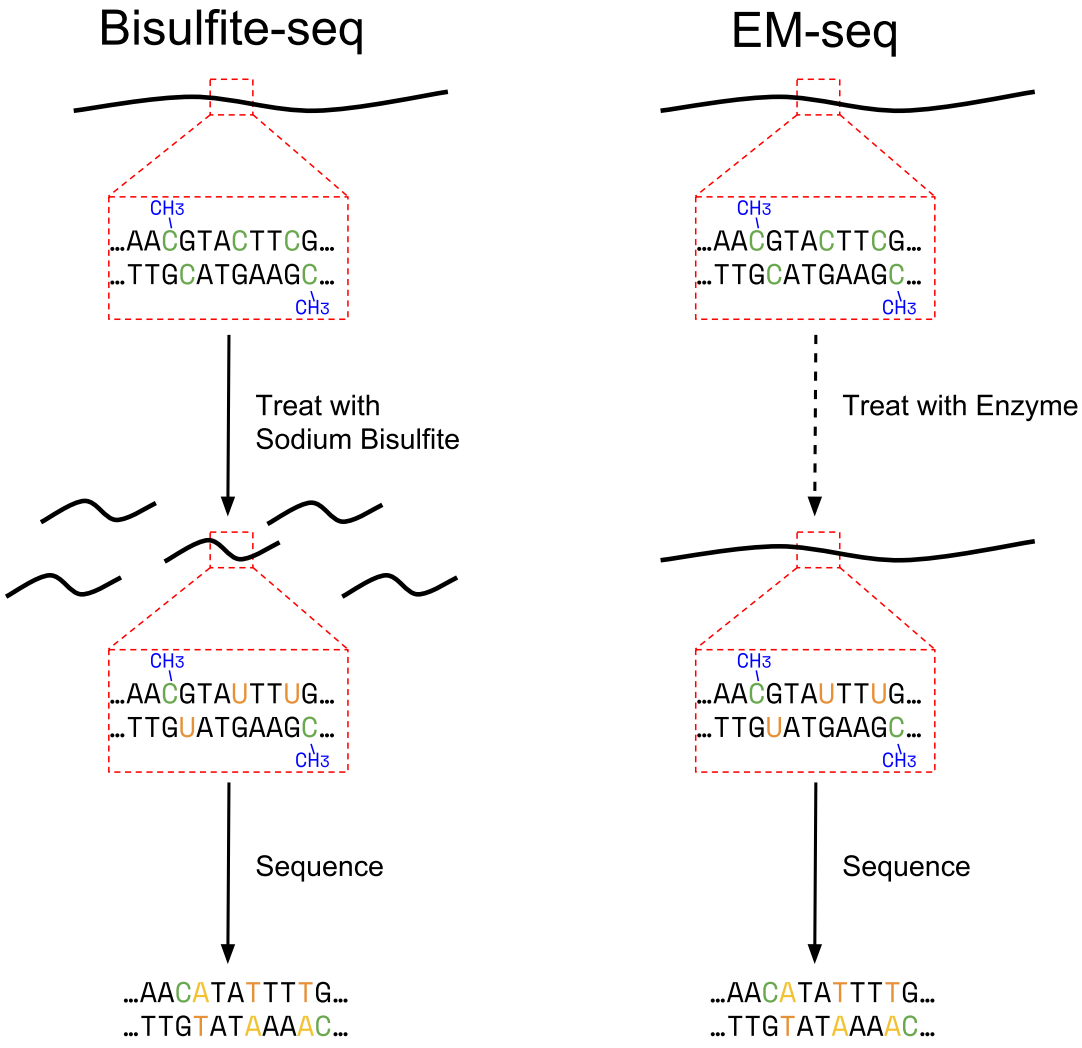

Figure 1: Overview of Methylation Sequencing Methods. <todo: Treat with Enzyme -> Tet2-> APOBEC

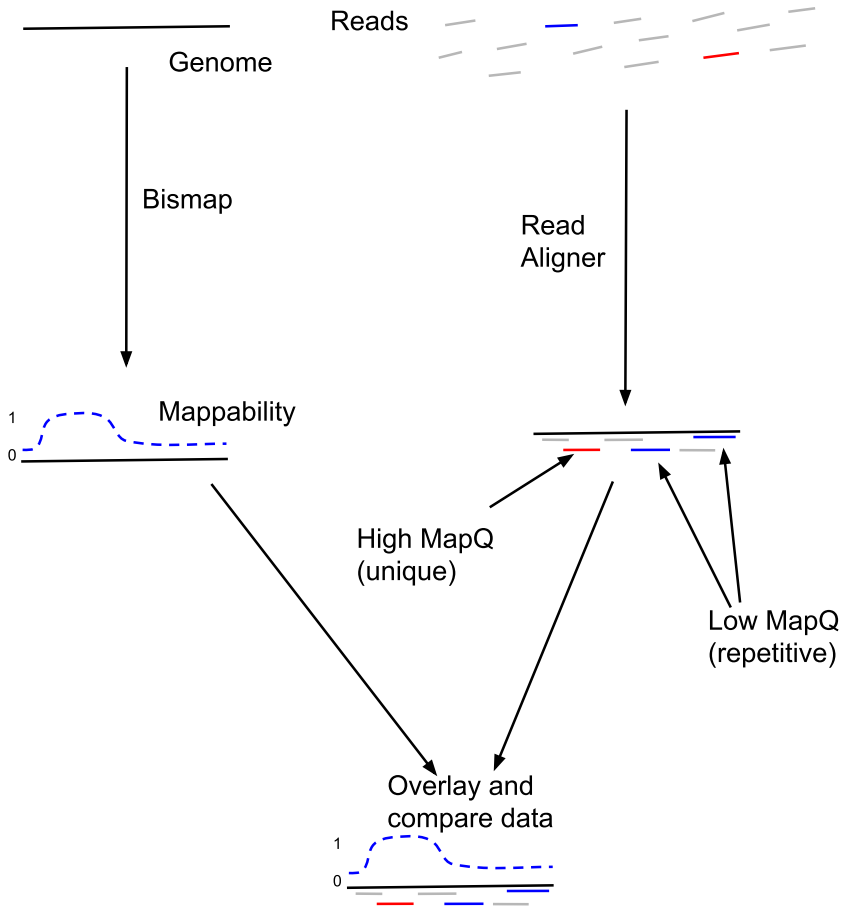

Figure 2: Experimental Overview

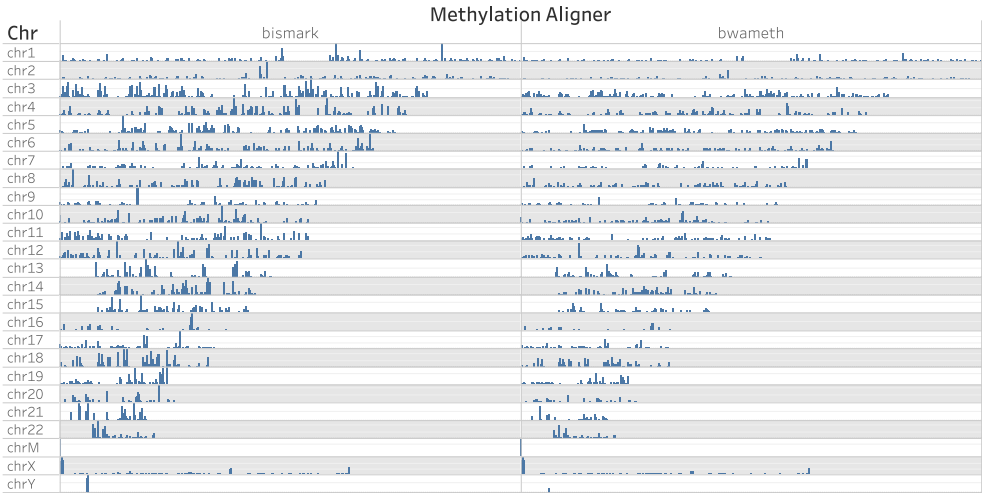

Figure 3: <WRONG CAPTION!!!!!> Example of MapQ Issue. On the Cov Difference track, red means that Bismark has higher coverage and blue means bwa-meth has higher coverage.

case, k=100) in the genome 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 and therefore produces data which is applicable in the context of methylation sequencing. Reads entirely contained within a region of low mappability should not have high MapQ values due to their repetitiveness, however we observed many high-MapQ reads in regions with very low or zero Bismap mappability (see Figure 4).

While low MapQ can indicate repetitiveness, 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 underfiltering. Cs in short, unique regions would be kept (underfiltered) even if the surrounding DNA is repetitive (see Figure 5). However, this does not happen often. Overfiltering of Cs in short repetitive regions is a larger problem though. 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 in GRCh38) but overfiltering is much more common (y% of Cs in GRCh38) <need to fill in data here> (see Figure 6). In cases where reads from multiple low-mappability regions are placed onto one region (creating a coverage spike), filtering by coverage could possibly also remove the problematic region, but in cases where reads are simply mixed among multiple low-mappability regions (not creating a coverage spike), coverage filtering would not be effective

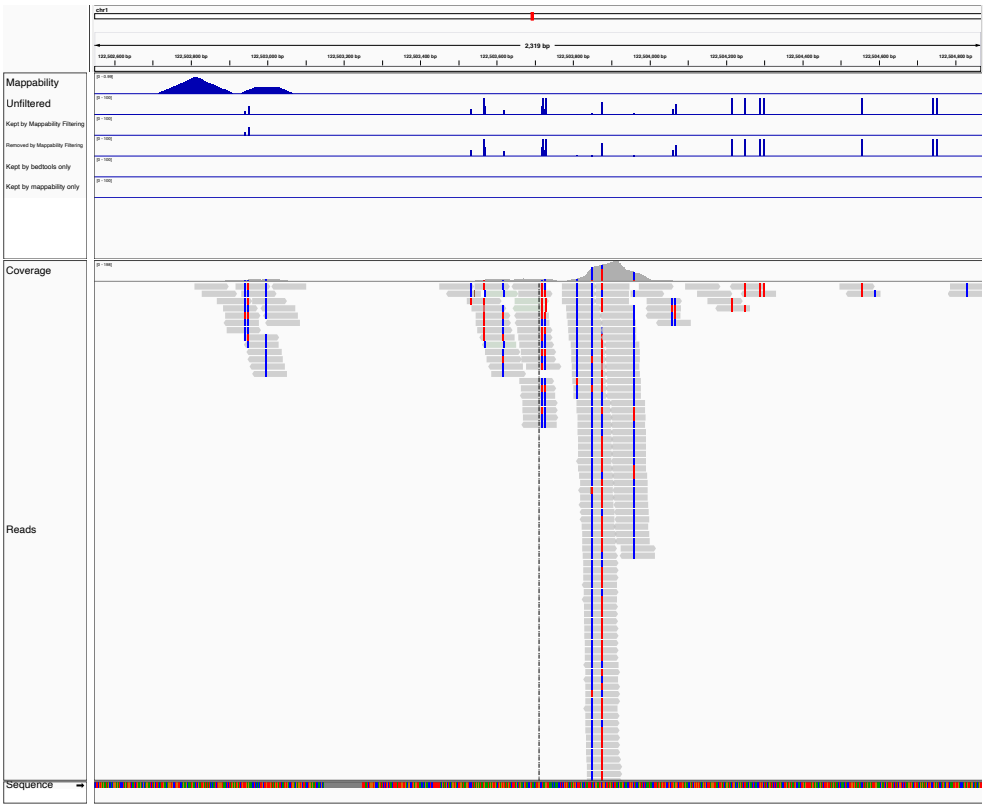

Figure 4: An example of how MapQ filtering does not remove all poorly mapped reads.

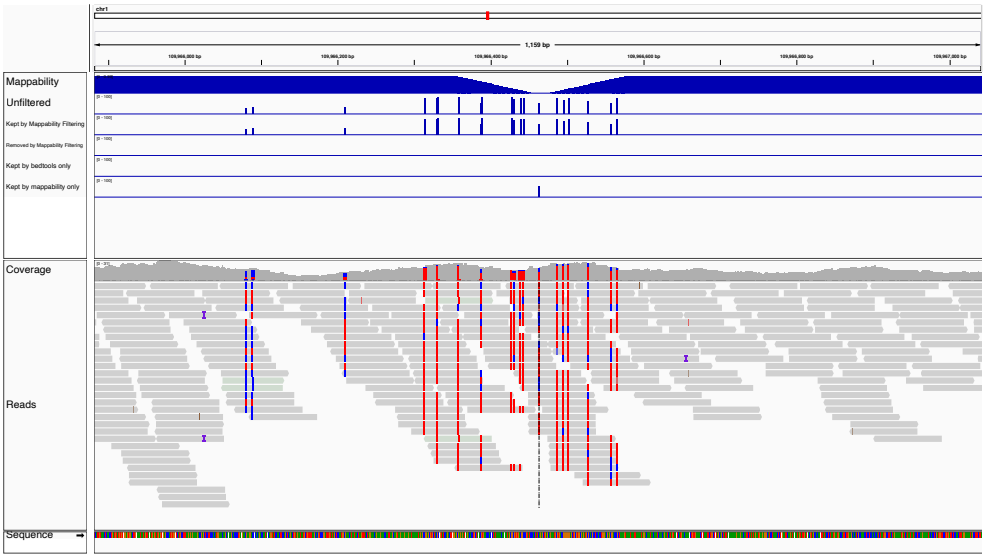

Figure 5: An example of a C incorrectly discarded by coordinate filtering.

at removing the affected regions.

To reliably filter out 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 alignment filtering approach precisely eliminates only those reads in repetitive regions and does not incur a significant cost in terms of execution speed (10:14 min with the patch 9:55 min without, when run with 20 threads on the Grch38 reference genome). (see computational methods for details)

To focus on the incorrect alignments and methylation calls that have biological and medical significance, we examined methylation calls in GENCODE genes [7] that contain variants listed in the ClinVar [13] database of disease-associated variants. Using our alignment filtering approach we successfully avoided calling 15,539 Cs in these important regions in a 50ngenzymatic sequencing sample which have insufficient mappability to support methylation calling. Briefly, to determine that we avoided calling a region has insufficient mappability to support methylation calling, we intersected 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).

For comparison, we also examined the Bismark methylation aligner [12] in the same manner. We observed that compared to bwa-meth, Bismark produced more high MapQ reads in repetitive regions (see Figure 3). In addition, Bismark and bwa-meth use different systems to define values for MapQ. Bismark (as a result of using Bowtie2 for alignment) reports MapQ based on number of reference mismatches, producing values between 0 and 42 [1, 9]. Bwa-meth (using BWA-MEM for alignments) follows the SAM specification in estimating

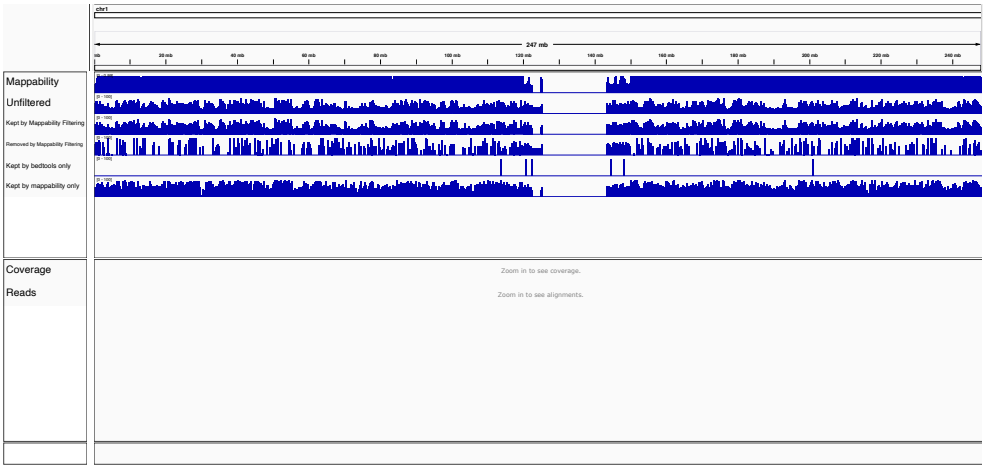

Figure 6: An overview of the occurrences of these scenarios in grch38 chromosome 1.

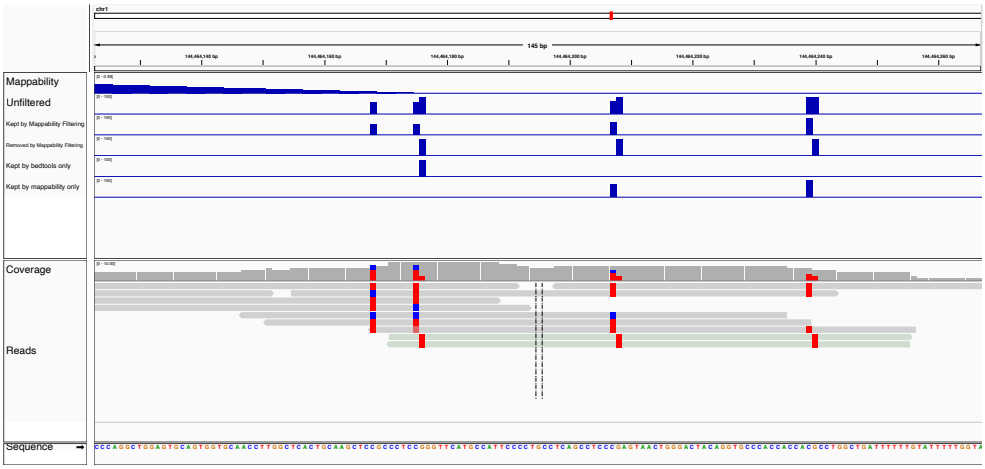

Figure 7: An example of a C incorrectly kept by coordinate filtering.

probabilities in the MapQ field.

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

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 bwa-meth 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 (I have 2 of these from my talk)

todo: find/examine relevant human (GEO?) dataset (maybe, depends on time/progress) (sequencing, not chip) (ideally, clinically relevant EM-seq)

todo: T2T data (repeat analysis on this and compare effect)

Materials and Methods

DNA Methylation Sequencing

Materials

Libraries were prepared from 10, 50, and 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 <is there a specific source for this, e.g. a company and/or product name?>, 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 <product number?> before being Bisulfite converted according to <protocol>. <anything about the EM-seq samples?>

Sequencing

Libraries were pooled and sequenced with diverse libraries (~10%) on 2 flow cells <is this still correct with the 2 other masses added?> of an Illumina Novaseq 6000 [8] using the S2 chemistry<citation>. We acquired 1.55 billion 99bp paired-end reads for the enzymatic sequencing method and 1.60 billion paired-end reads for the Bisulfite converted libraries <what mass is this for? need counts for all masses unless they're identical>.

Computational Methods

In order to run the analysis, the following data and tools were used:

The GRCh38.p11 analysis set (hereafter referred to as "grch38") supplemented with phage T4, phage lambda, phage Xp12, and pUC19 contigs was used throughout [4]. A VCF file of disease-associated variant sites was downloaded from ClinVar (see supplemental data for file date) and a GFF file of the GENCODE v31 gene annotations were used. The mappability data was the 100bp multi-read bigWig file downloaded from Bismap (see supplemental materials for link). A detailed diagram of the analysis pipeline is found in Figures 9 and 10. Tools used include bedtools [17], samtools [14], GNU awk, MethylDackel, bigWigToBedGraph [11], BEDOPS [15], GNU sort, GNU head, bwa-meth, and Bismark. GNU sort and head are required since some of the functionality needed (specifically, the ability to parallelize sorting and the ability 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.

The sequencing reads were aligned using the bwa-meth aligner using default options. Methylation calling was performed on the resulting BAMs using MethylDackel v0.5.2. MethylDackel uses a default value of 10 as a minimum MapQ which should include mostly single locus reads, 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 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 mappability 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 an FAI index of the grch38 reference genome FASTA 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

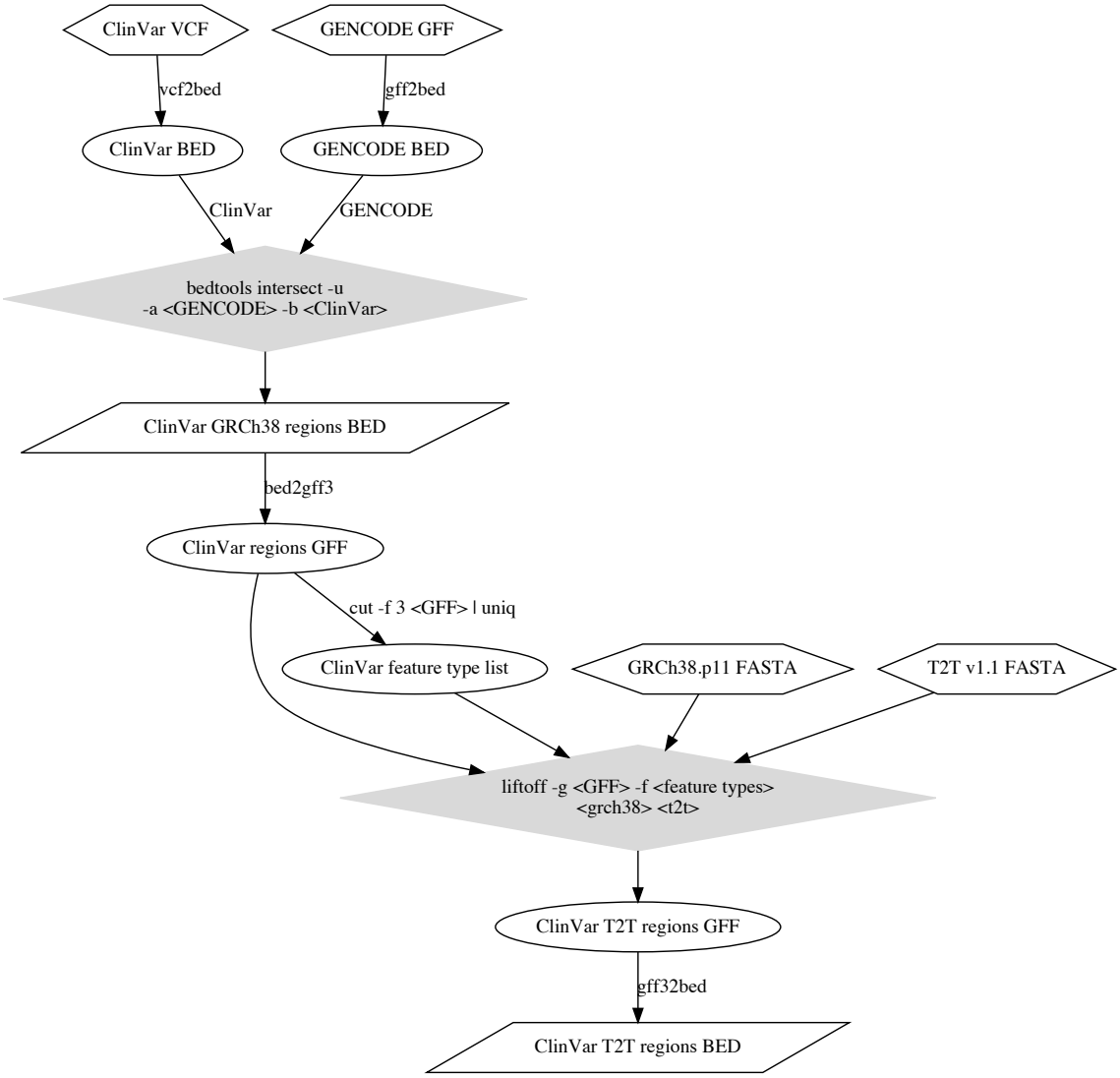

Figure 8: Detailed description of preprocesing steps for ClinVar and GENCODE data. <todo: add legend with figure features, boxes, ovals, edge types, etc. parts A and B with >

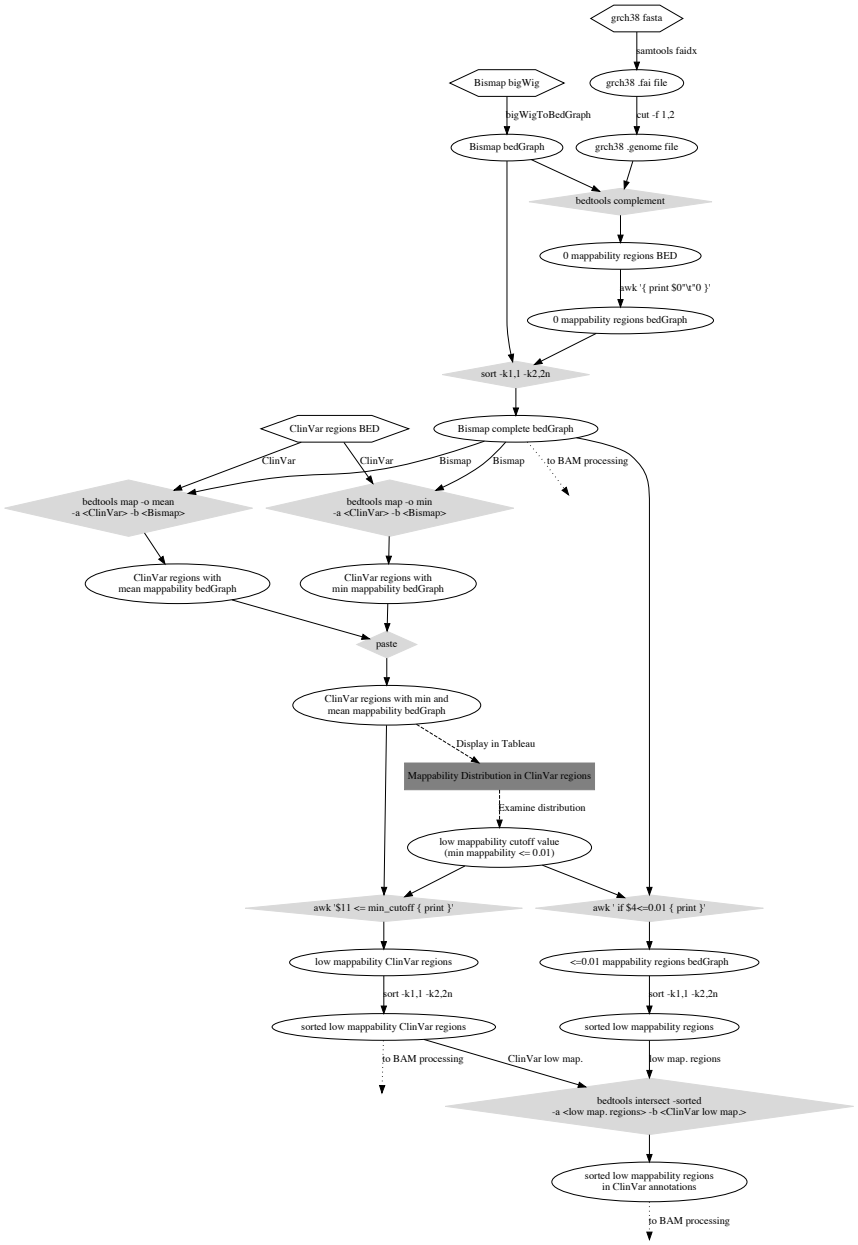

Figure 9: Detailed description of preprocesing steps in data analysis. <todo: add legend with figure features, boxes, ovals, edge types, etc. parts A and B with >

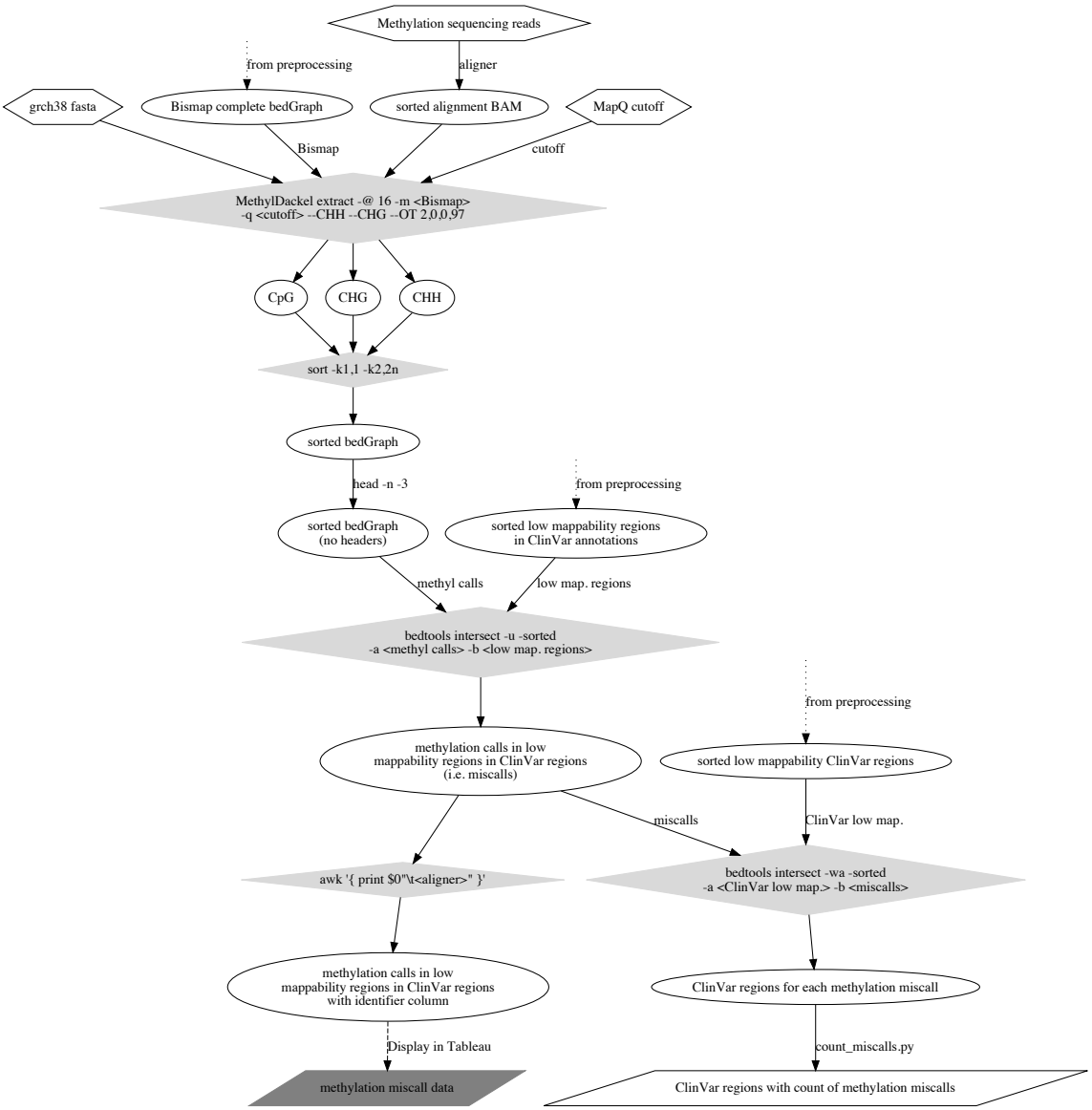

Figure 10: Detailed description of BAM processing steps in data analysis. <todo: add legend with figure features, boxes, ovals, edge types, etc. parts A and B with >

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. <what did I do for partial overlaps here??>

Alignments of 2x99-bp paired-end enzymatic sequencing and whole genome bisulfite reads were processed using MethylDackel (with and without the custom patch) using a minimum MapQ cutoff of 10 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, but was not filtered out by basic MapQ 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, which writes a copy of the ClinVar region to the output file for each miscall in the region, in order that this file would contain multiple copies of each region, one per miscall in the region. These duplicates were then used to count miscalls by feeding the data to a custom Python script (found in the Nextflow script in the supplemental materials) which counted and combined the duplicates, producing a list of all ClinVar regions with miscalls and how many miscalls are in each region.

Since this analysis was run for three different input masses and two sequencing protocols, the miscalls files were also processed through awk to add a field specifying which input mass, sequencing protocol, and MethylDackel filtering setting were used. The same field was added to the lists of all ClinVar regions with miscalls and counts described previously. All files were examined and compared in Tableau®.

To compare the behavior of Bismark with bwa-meth, this analysis was re-run with the Bismark aligner <with what alignment options>. The name of the aligner was added to the field specifying the input mass, protocol, and MethylDackel filtering settings present in the miscalls files and the lists of all ClinVar regions with miscalls and counts.

<still getting data>

To compare the need for and the effect of mappability filtering on the more-complete T2T v1.1 reference genome <citation> (hereafter referred to as "T2T"), the analysis was re-run using the T2T reference in place of grch38 (supplemented with phage T4, phage lambda, phage Xp12, and pUC19 just as with grch38). To use the ClinVar and GENCODE data, which were respectively originally obtained as VCF and GFF files containing genomic coordinates in grch38, with the T2T reference, the positions were transferred over to the T2T reference using Liftoff <cite>. Specifically, the ClinVar regions BED (produced from the ClinVar VCF and the GENCODE GFF) was converted to GFF3 with a custom tool <add to supplemental> because Liftoff requires GFF files. This file was lifted over to the T2T reference, and then converted back to BED using a second custom tool <supplemental>. This converted file was used in place of the ClinVar regions BED for analysis with the T2T reference. As there was no

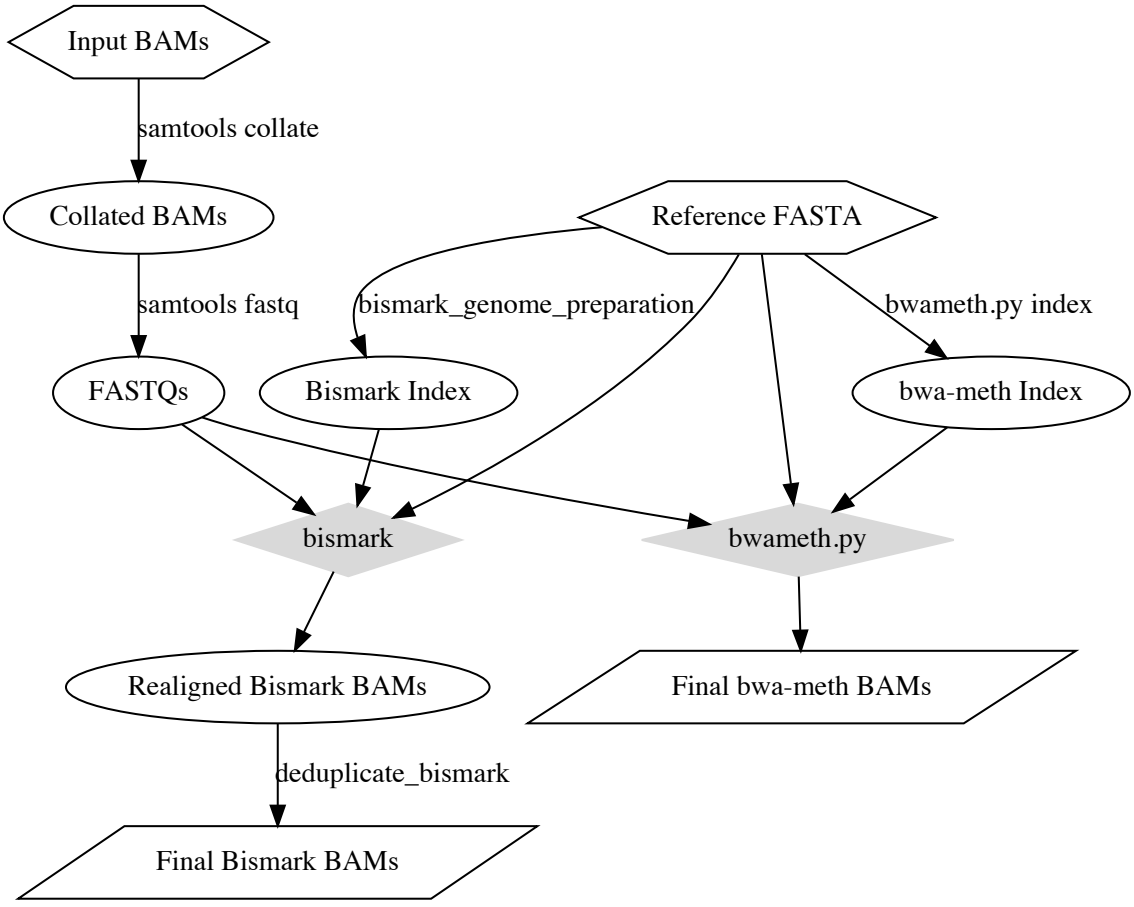

Figure 11: Detailed description of BAM realignment steps. <todo: add legend with figure features, boxes, ovals, edge types, etc. parts A and B with >

Bismap mappability file for the T2T reference, we downloaded the Bismap tool and generated a k=100 multi-read Bismap mappability dataset for this reference genome. <expand on how Bismap run was done? or leave it at that> <finish this analysis and write up how it was done>

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 bwa-meth for alignment and

MethylDackel with MapQ > 20 for methylation calling <where is the test of MapQ 10 vs 20? or was >10 intended?>. 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 (0.5.0 or newer, we performed this analysis with version 0.5.2), GNU sort, GNU head, bedtools, BEDOPS, paste, awk, and Python 3.

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 0.5.0.

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

The ClinVar VCF was the July 22, 2019 version of ClinVar's variants VCF, with a file name of clinvar_20190722.vcf.gz

References

[1] QC Fail Sequencing Â» MAPQ values are really useful but their implementation is a mess.

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

[3] 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].

[4] 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.

[5] 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].

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

[7] 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].

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

[9] Johnurbangenome. Biofinysics: How does bowtie2 assign MAPQ scores?, May 2014.

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

[11] 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/.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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