Correcting Methylation Calls in Low-Mappability Regions

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

August 13, 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 and removes many unsupportable methylation calls. 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 [29], 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].

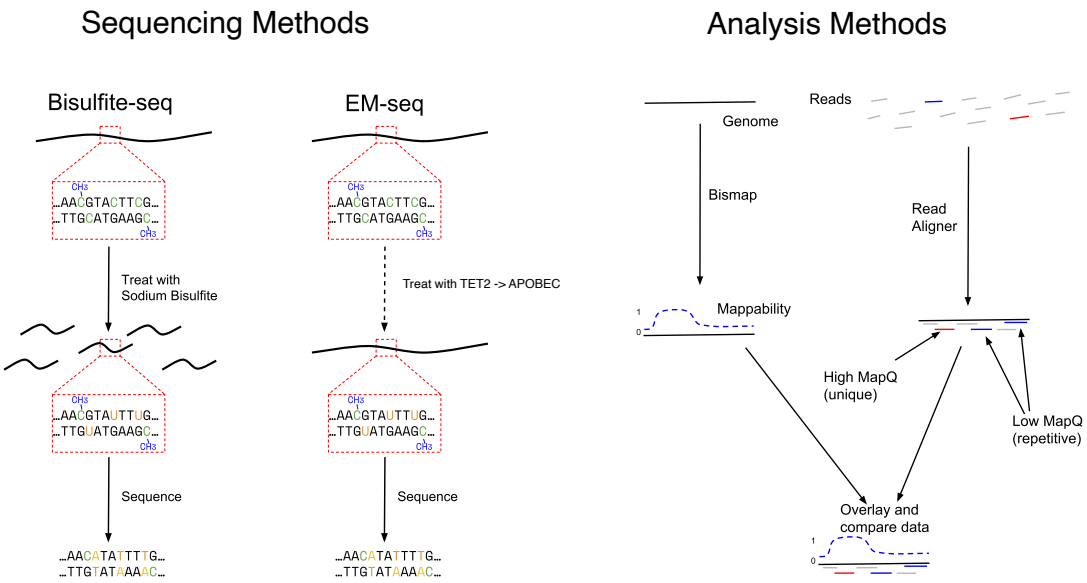

Figure 1: Overview of Methylation Sequencing and Data Analysis Methods.

It is also possible to use an enzymatic method, EM-seq, 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 [31].

Whichever method is used to deaminate cytosines, sequence data is typically aligned to a reference genome using a methylation-aware aligner [6], 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 [28] or bismark_methylation_extractor [15], which will use the resulting read alignments to determine the methylation status of particular cytosines (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 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 1). According to the SAM specification [8], 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 [24], 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 [12], 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 (ranging from 0 to 1 in increments of 0.01) for every base in the GRCh38 reference. Each individual K-mer is treated as either mappable or not (binary), and the overall mappability score is derived from how many different K-mers are mappable or not. Bismap takes the effect of C->T conversion into account and therefore produces data which is applicable in the context of methylation sequencing [12]. 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 3).

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 situation was not commonly observed. 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 6). In practice, underfiltering is rare (a median of $7.756 * 10^{-5}\%$ of called Cs in GRCh38 when calling using EM-seq reads aligned with bwa-meth) but overfiltering is much more common (a median of1.054% of called Cs in GRCh38 when calling

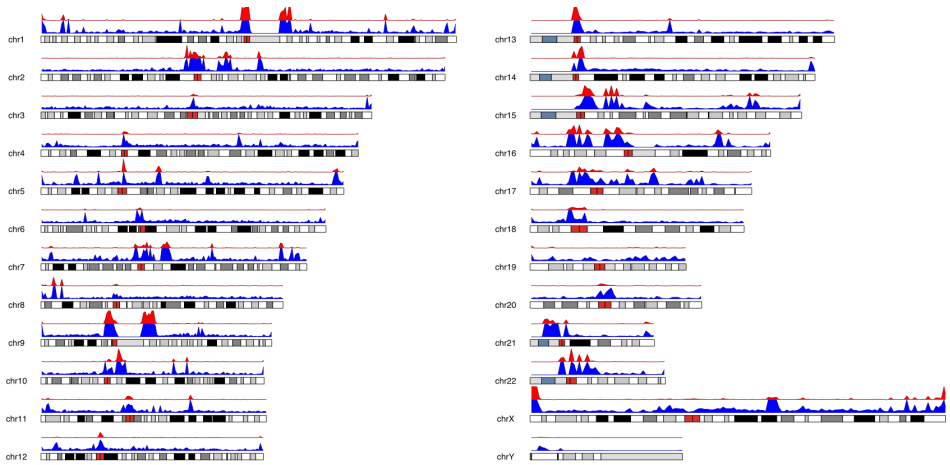

Figure 2: Cs in low mappability regions, as well as thosefiltered out by per-read filtering that still passed MapQ filtering, are represented across all chromosomes. However, per-read filtering can remove unsupported calls while removing far less Cs than a naïve per-C filtering approach would require. In this figure, the horizontal banded rectangles are the chromosomes, the blue track above them is all calls in low-mappability rgeions, and the red track is miscalls removed by per-read filtering.

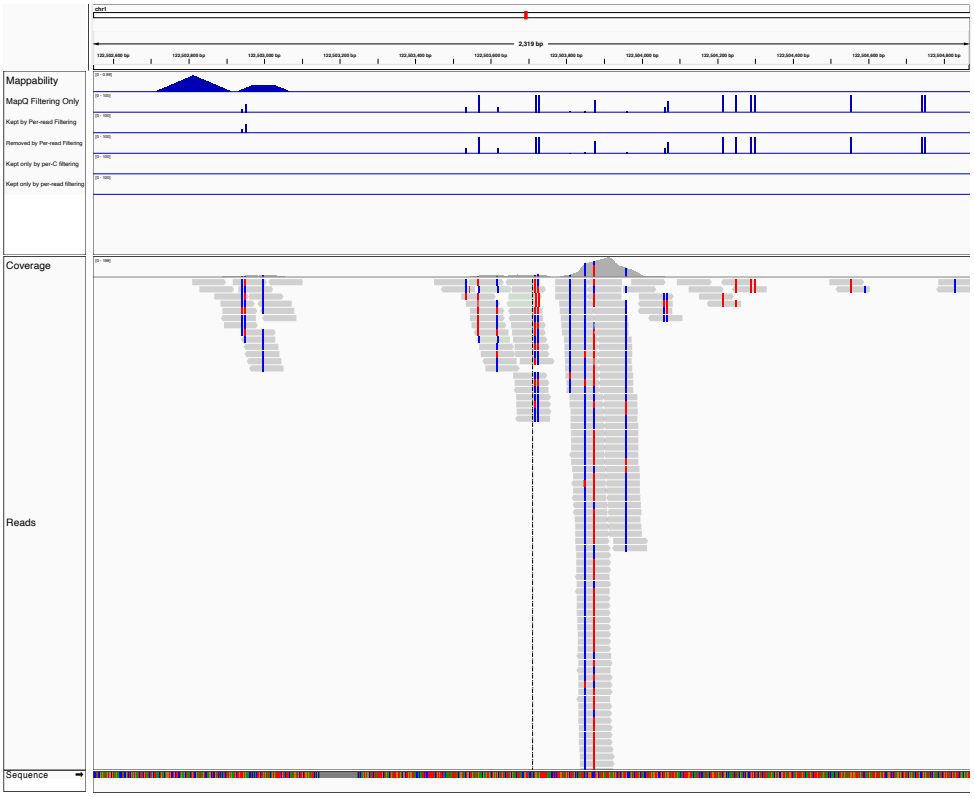

Figure 3: MapQ filtering does not remove all poorly mapped reads. This is an example of a coverage spike in a low-mappability region: clearly not sufficiently mappable, but still called with just MapQ filtering.

using EM-seq reads aligned with bwa-meth) (see Figure 4). Although even about 1% of called Cs might seem small, the total number of called Cs without filtering is large enough that that 1.054% is on the order of 12.1 million Cs, while $7.756 * 10^{-5}$% of Cs is closer to 900 Cs. 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 at removing the affected regions. In addition, if some of the reads used to call a C are uniquely mappable, but others are not, per-read filtering can remove only those reads that are not uniquely mappable, rather than being forced to discard the entire locus like per-C filtering does.

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 per-read 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 <update these numbers?>). (see computational methods for details)

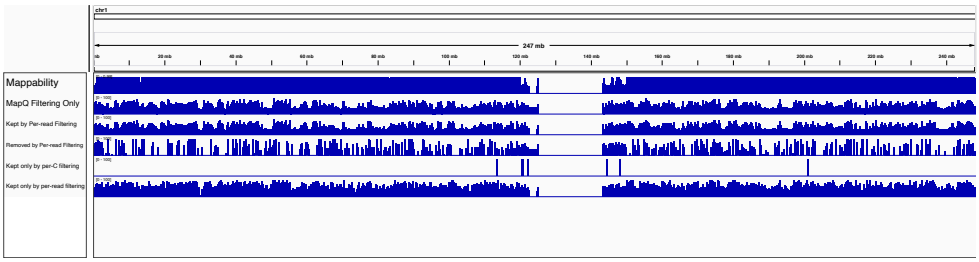

Figure 4: An overview of the occurrences of these scenarios in GRCh38 chromosome 1. The six tracks show the effect of per-C and per-read mappability filtering on the methylation calls on this chromosome. "Mappability" is the Bismap mappability for the chromosome. "MapQ filtering only" is the calls after MapQ filtering only (no per-C or per-read mappability filtering). "Kept by Per-read Filtering" shows the methylation calls kept after per-read mappability filtering. "Removed by Per-read Filtering" shows those calls removed entirely by per-read filtering for being insufficiently mappable. "Kept only by per-C filtering" shows the underfiltered Cs, that is, those that are not sufficiently mappable according to per-read filtering, but are retained by per-C filtering. "Kept only by per-read filtering" shows the overfiltered Cs, that is, those that are sufficiently mappable according to per-read filtering, but are removed by per-C filtering.

To focus on the incorrect alignments and methylation calls that have biological and medical significance, we examined methylation calls in and just upstream of GENCODE genes [9] that contain variants listed in the ClinVar

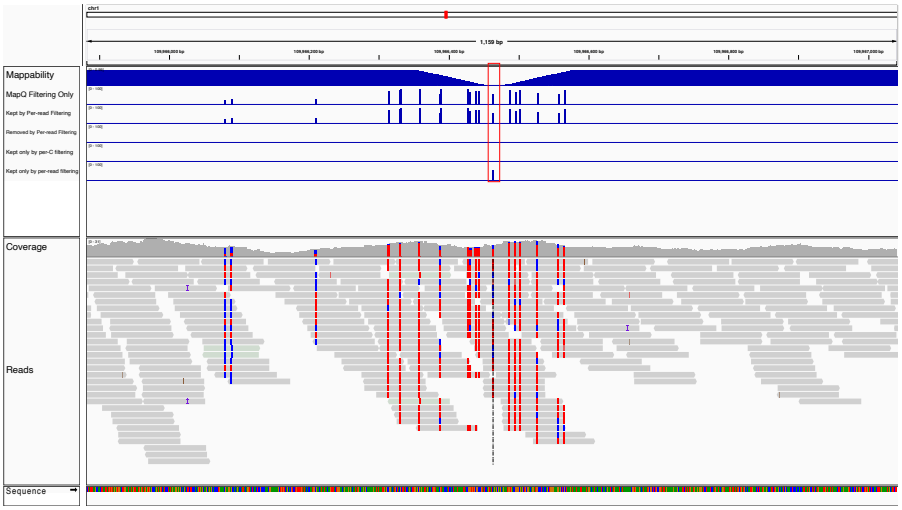

Figure 5: An example of a C incorrectly discarded by coordinate filtering (circled). The C is in a small mappability dip, causing it to be discarded by bedtools filtering, while per-read filtering correctly retains the C, as it can be called using reads extending into the high-mappability regions on either side.

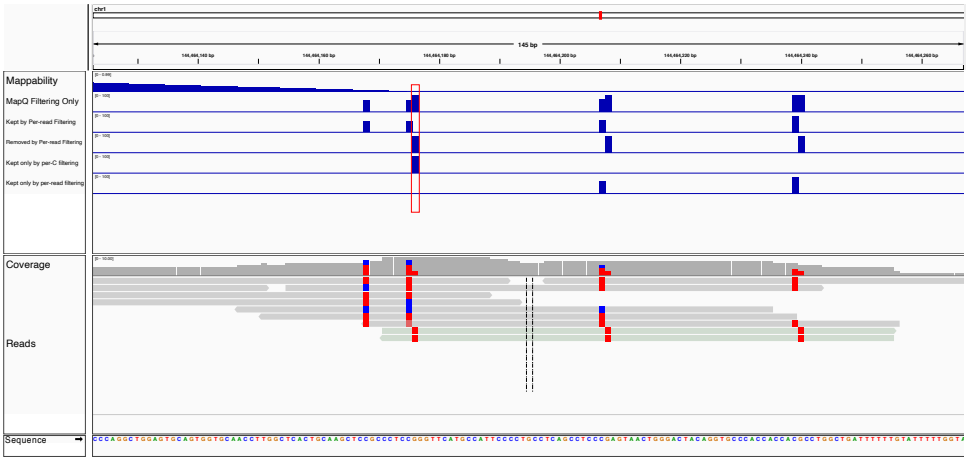

Figure 6: An example of a C incorrectly kept by coordinate filtering (circled). As the C is itself just inside the edge of a mappable region, bedtools retains it, even though it is only supported by two insufficiently mappable reads. Our per-read filtering removes this sort of C, as there is not enough data here to confidently call it.

[17] database of disease-associated variants. Using our alignment filtering approach we successfully avoided calling 275,092Cs in these important regions in a 50ng EM-seq sample which have insufficient mappability to support methylation calling. Briefly, to determine that we avoided calling methylation for a C for which there is not a sufficiently mappable read, we took the list of unfiltered Cs for the regions we were interested in and removed from it all Cs we kept after per-read filtering. The result was a list of every C that was filtered out (see Figure **??**).

Filtering Effect

We compared the effect of per-read mappability filtering on number of C's called, as well as number of C's called in low-mappability regions (as defined by per-C filtering), on EM-seq samples of 3 sample masses aligned with the bwa-meth aligner. We saw a significant positive filtering effect on all 3 sample masses, with a mean of 7.67 million Cs filtered out by read-based filtering in each sample (see Figure 7). The miscalled Cs were scattered across the genome, as shown in Figure 2.

We also examined how many mixed methylation ($0\% < \%\text{Methyl} < 100\%$) Cs were resolved to either 0% or 100% methylation, and found that approximately 60,000-90,000 methylation calls in the EM-seq data that were previously mixed were resolved by per-read filtering to either fully unmethylated or fully methylated (see Figure 8), showing how unlike per-C filtering, per-read filtering can clean up methylation calls without always having to entirely delete them.

Effect on Potentially Significant Genes

We also assessed the biological significance of these miscalls. To measure this, we used genes in GENCODE containing variants in ClinVar (henceforth "ClinVar genes") as a metric for biological significance. To focus on genes that may contain miscalls, we filtered the list to only include those genes which had at least one region, either in the gene or in the 2kb upstream of the gene (in order to capture regulatory regions), where the mappability is too low to call methylation (as without this, there will be no miscalls in the region). Examination of this data in IGV[27] revealed that there were miscalls in or upstream of 1447-1586 unique genes in the EM-seq/bwa-meth data, with a median number of 64 miscalls per gene region, indicating that indeed this may affect biologically significant methylation data. It is evident that the sites of most methylation miscalls fall into two broad types. First, there are those regions that simply have a mappability of 0 for a long stretch of sequence, resulting in all Cs that were called in this region being miscalls (as there are no mappable regions to anchor reads in). Second, there are those regions at the border between high mappability and low mappability. These regions are where overfiltered (and underfiltered) Cs (by per-C filtering, as opposed to our per-read filtering) are found: overfiltered Cs occur when a C is in the low mappability region, but there is a mappable read anchored in the high mappability region that allows it to be called. Underfiltered

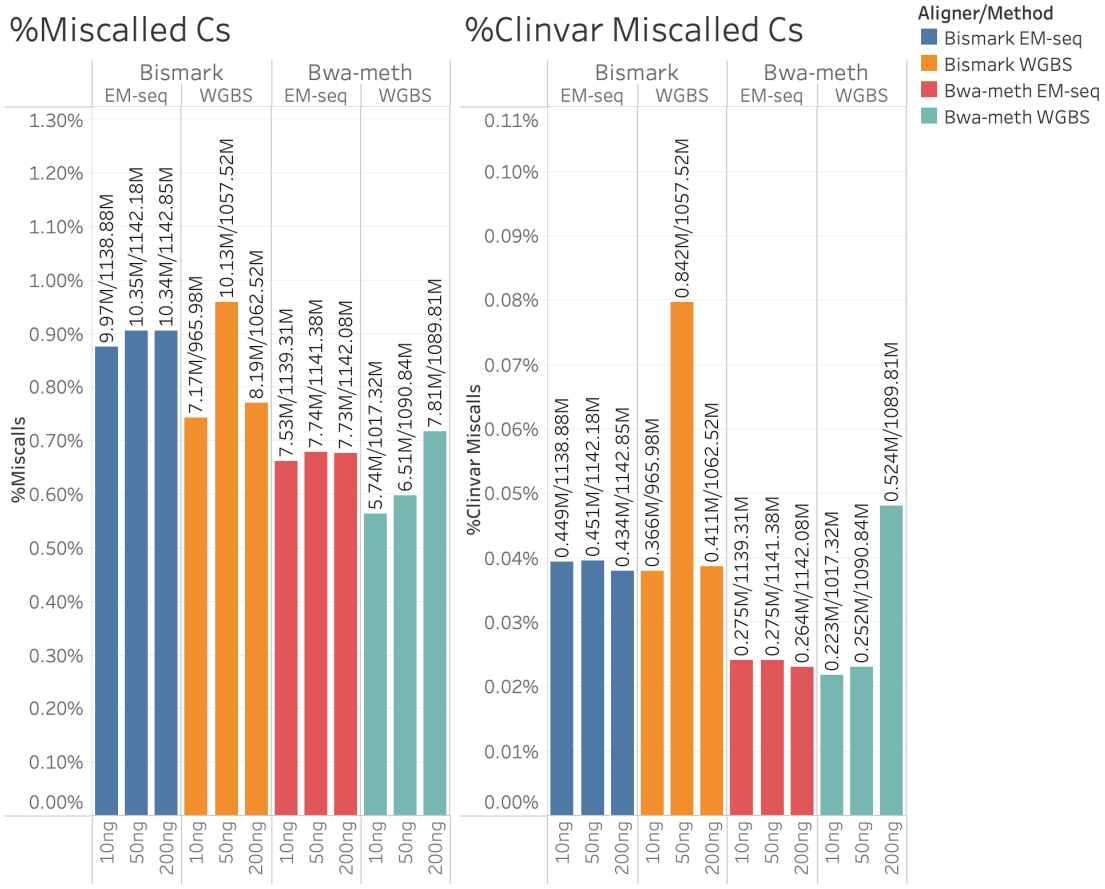

Figure 7: Left: Percents of miscalled Cs in the 10, 50, and 200ng EM-seq and WGBS samples, aligned with bwa-meth and Bismark, where a "miscall" is a methylation call filtered out entirely by per-read filtering. The vertical axis is the percent of miscalled Cs in the sample. Right: Percents of miscalls within ClinVar genes in the same samples. The vertical axis is the percent of ClinVar miscalled Cs in the sample. In both figures, the colors define which sequencing method and aligner were used, as detailed in the legend.

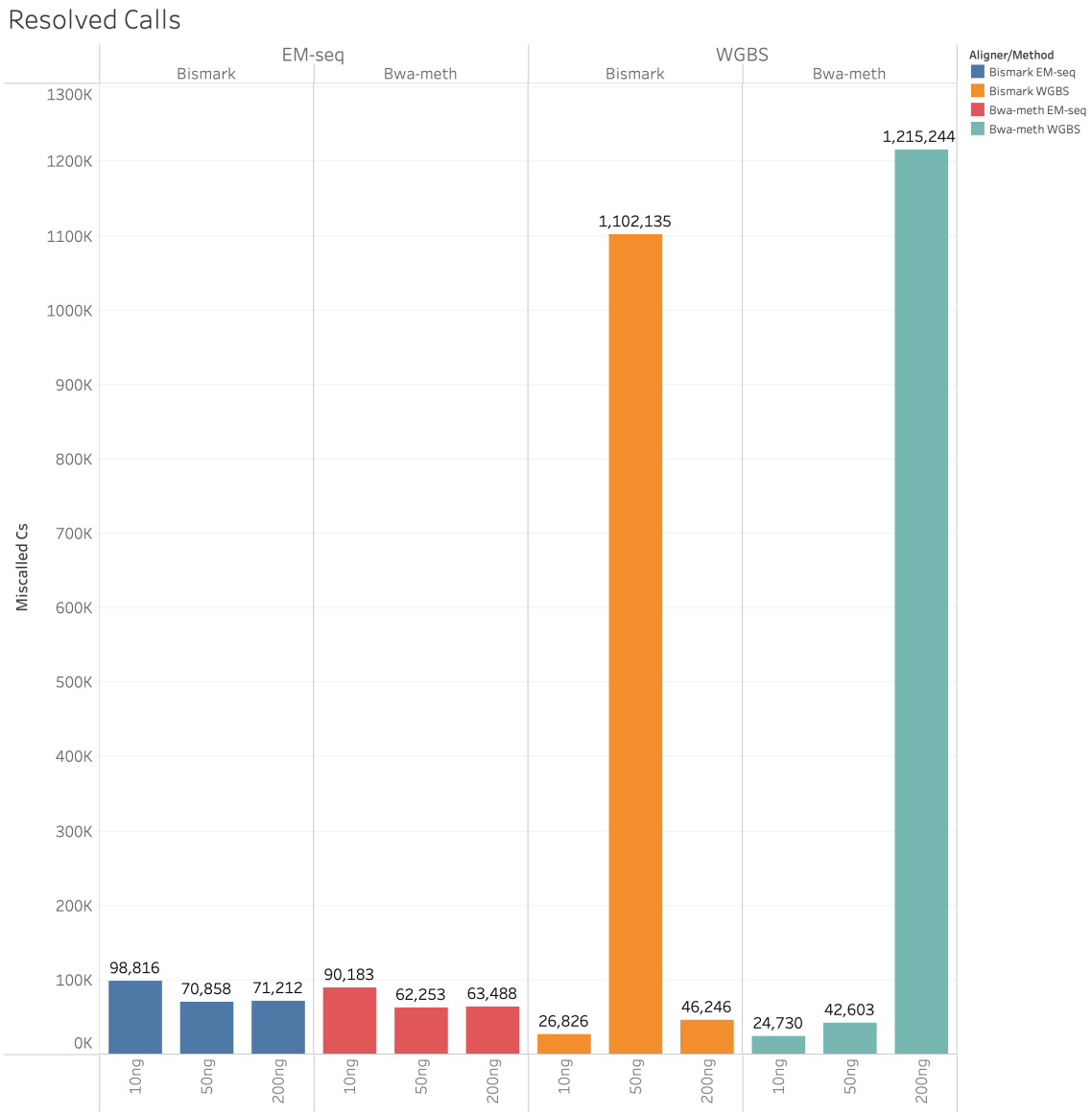

Figure 8: Counts of resolved calls in the 10, 50, and 200ng EM-seq and WGBS samples, aligned with bwa-meth and Bismark, where a "resolved call" is a methylation call that is mixed in the unfiltered data, but either 0% or 100% in the filtered data. The vertical axis is the number of resolved calls in the sample. The colors define which sequencing method and aligner were used, as detailed in the legend.

C's occur when a C is just inside a mappable region, but there is no sufficiently mappable read to call it. An example of each (from the 200ng EM-seq sample aligned to GRCh38 using bwa-meth) is provided in Figures 9and 10 respectively.

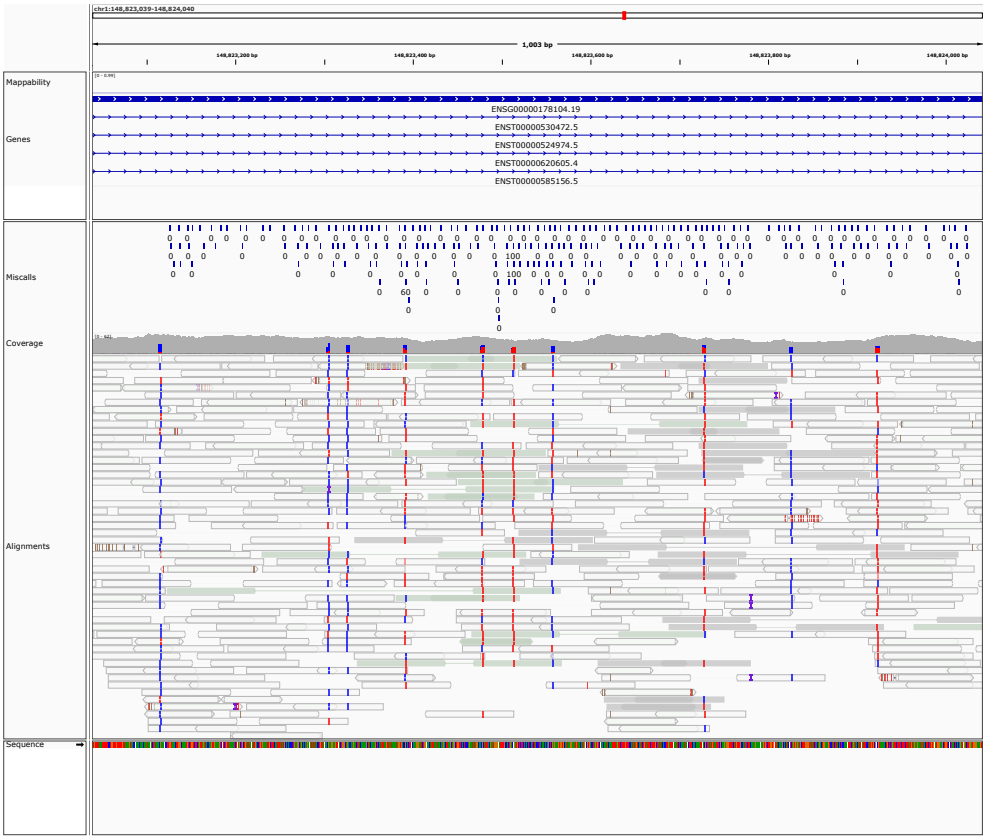

Figure 9: A view of a cluster of miscalls in a large low-mappability region. The gene, ENSG00000178104.19, is known as PDE4DIP. The reads are shown paired, with a line connecting reads 1 and 2 in a pair, with CpG methylation sites highlighted. Miscalls are shown in the track above the coverage graph. Note how every CpG call (as well as calls in other contexts) are miscalled in this region, as there are no high-mappability regions to anchor reads in.

EM-seq vs. WGBS

Due to the reduced DNA damage and resulting longer possible read lengths that occur in EM-seq as opposed to WGBS (whole genome bisulfite sequencing), we compared two sets of sequencing runs, each with the same 3 sample masses, to see if there is a difference between the two aligners with regard to the number of miscalls, as it is possible longer reads could allow for more low-mappability Cs

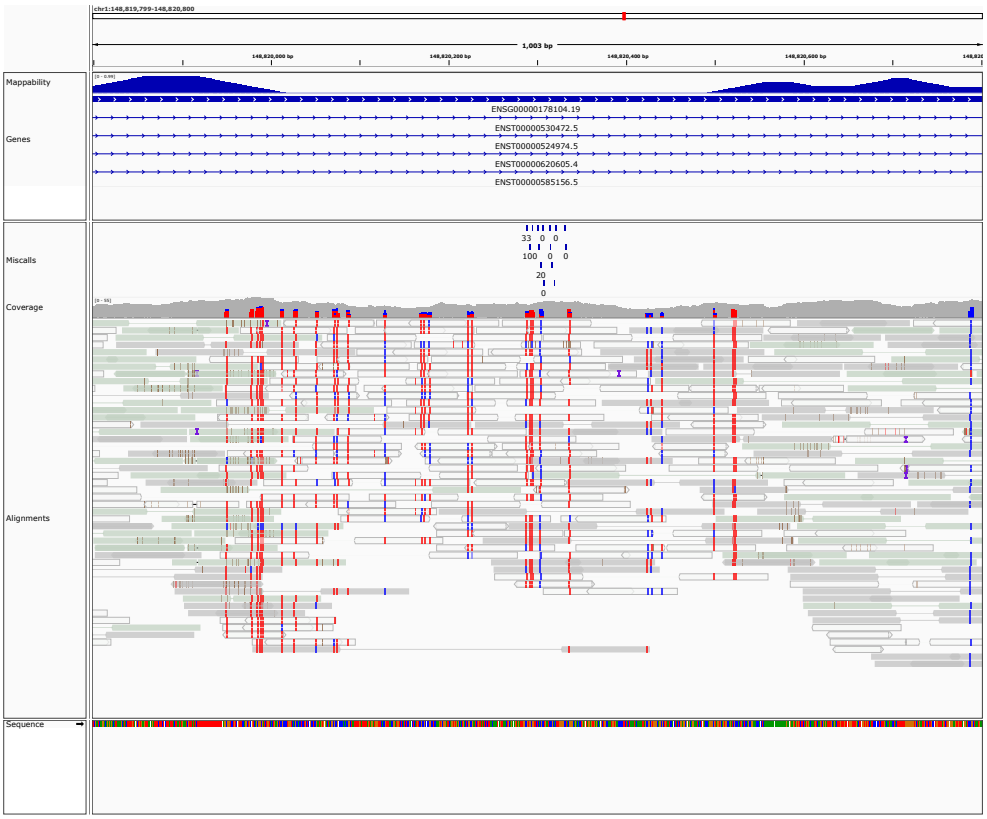

Figure 10: A view of a cluster of miscalls near the boundary between high and low mappability regions. The gene, ENSG00000178104.19, is known as PDE4DIP. The read alignments are shown paired, with a line connecting reads 1 and 2 in a pair, and highlighted for CpG methylation. Note the methylation calls in the low-mappability region that are not miscalls, as they can be called using read pairs extending into the high-mappability region.

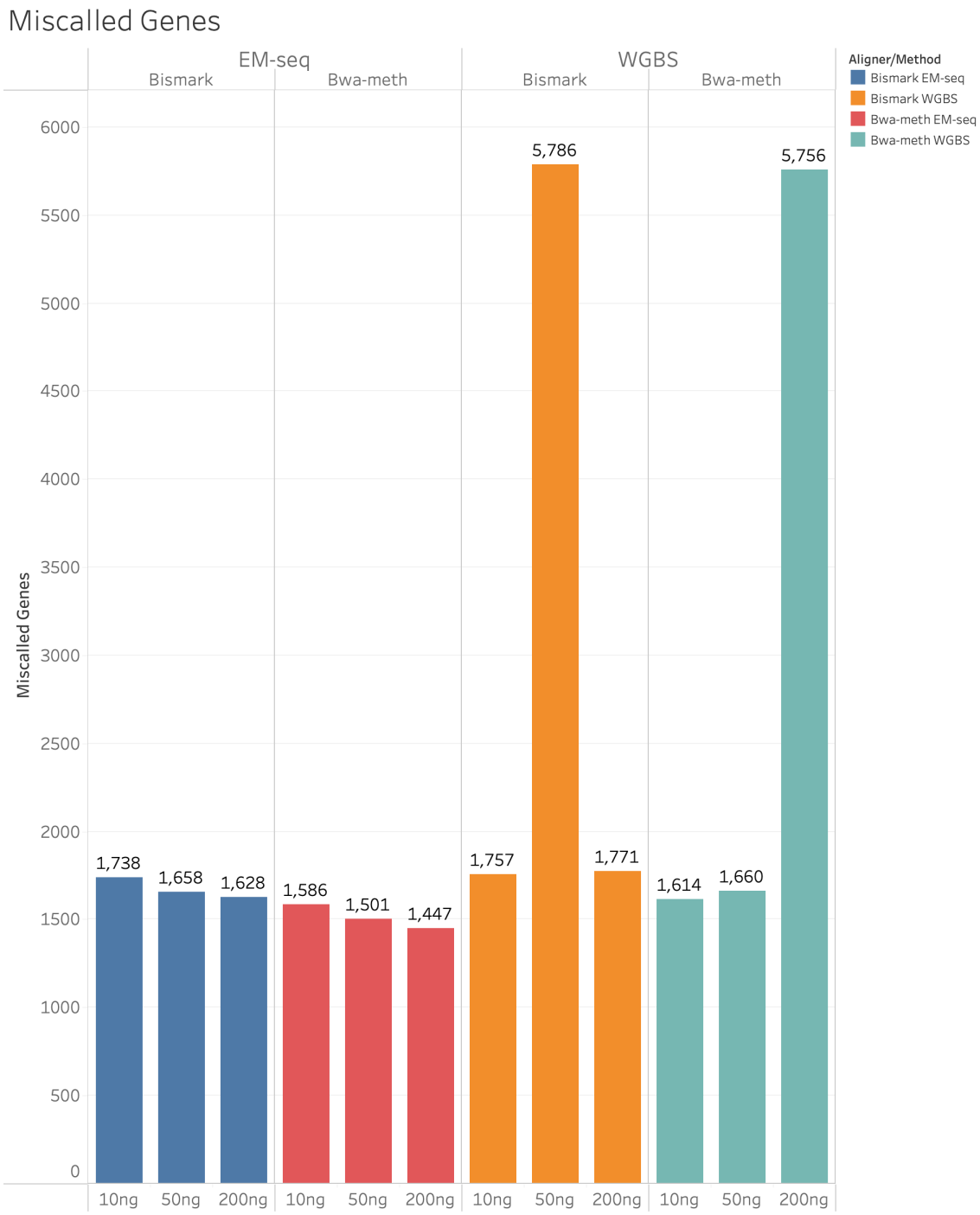

Figure 11: Counts of genes containing miscalls in the 10, 50, and 200ng EM-seq and WGBS samples, aligned with bwa-meth and Bismark. The vertical axis is the number genes containing miscalls in the sample. The colors define which sequencing method and aligner were used, as detailed in the legend.

to be called using reads extending into higher-mappability regions. We found that EM-seq may have a marginally higher percentage of miscalled Cs at some input masses, but the effect is not very large (see Figure 7)

For miscalls in ClinVar genes, there is no real difference (discounting the two outliers, which will be discussed shortly) between the percent of miscalls with EM-seq as opposed to WGBS. However, there are two WGBS samples that are significant outliers across many of the analyses we performed: the 200ng WGBS sample aligned to GRCh38 with bwa-meth, and the 50ng WGBS sample aligned to GRCh38 with Bismark (see Figures 7, 8, and 11). <WHY THE OUTLIERS HAPPEN>

Bwa-meth vs. Bismark

For comparison, we also examined the Bismark methylation aligner [15] in the same manner. We observed that compared to bwa-meth, Bismark produced more high MapQ reads in repetitive regions (see Figure 2). 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 [25, 11]. Bwa-meth (using BWA-MEM for alignments) follows the SAM specification in estimating probabilities in the MapQ field. Overall, we observed a somewhat higher percent of miscalls with Bismark than we did with bwa-meth, as shown in Figure 7. This difference was distinct in both overall miscalls and miscalls in ClinVar genes.

GRCh38 vs. T2T

We also examined the effect of using the T2T reference [23] in place of GRCh38, since the T2T reference is much more complete than GRCh38, including repetitive regions such as centromeres, telomeres, and rDNA, which could play a role in where methylation is called and what areas of the genome have unique enough sequence to support calling. We found that, as we predicted, there were far fewer miscalls in the T2T data as compared to the GRCh38 data, with less than half the number of miscalls in T2T as in GRCh38, with total calls being roughly in the same range between the two reference sequences (see Figure 12). This indicates that it is indeed true that using the T2T reference can allow for better calling of methylation (as shown by the reduced percentage of miscalls).

todo: filtering effect em-seq vs wgbs (fig) for bwa-meth (in pipeline) [DONE]

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

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) (in pipeline, somewhat) [???]

todo: group Cs by magnitude 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 (in pipeline) [IN PROGRESS]

todo: examples of specific effects in ClinVar (I have 2 of these from my talk) [IN PROGRESS]

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) (in pipeline) [IN PROGRESS]

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 (isolated from Xp12 phage following the procedure in Y. J. Lee, P. R. Weigele, Detection of modified bases in bacteriophage genomic DNA. Methods Mol. Biol.2198, 53-66 (2021).[18])[16], lambda phage DNA (NEB #N3011), and a pUC19 plasmid (NEB #N3041) treated with M.SssI CpG Methyltransferase (NEB #M0226).

Whole Genome Bisulfite Libraries

The whole genome bisulfite (WGBS) libraries were prepared using the Ultra II DNA library prep kit (NEB E7645) before being Bisulfite converted using the Zymo EZ DNA Methylation-Gold bisulfite conversion kit according to the protocol described in Vaisvila et al. [31].

EM-seq Libraries

The EM-seq libraries were prepared using the NEBNext Ultra II DNA library prep kit (NEB E7645) (using the NEBNext EM-seq adapter), then EM-seq converted using TET2 and ABOBEC3A, as described in Vaisvila et al. [31].

Sequencing

Libraries were pooled and sequenced with diverse libraries (~10%) on 2 flow cells of an Illumina Novaseq 6000 [10] using the S2 chemistry. We acquired about 1.30 billion 99 bp paired-end reads for the EM-seq method and 1.27 billion paired-end reads for the Bisulfite converted libraries.

Computational Methods

The following data and tools were used for the analysis:

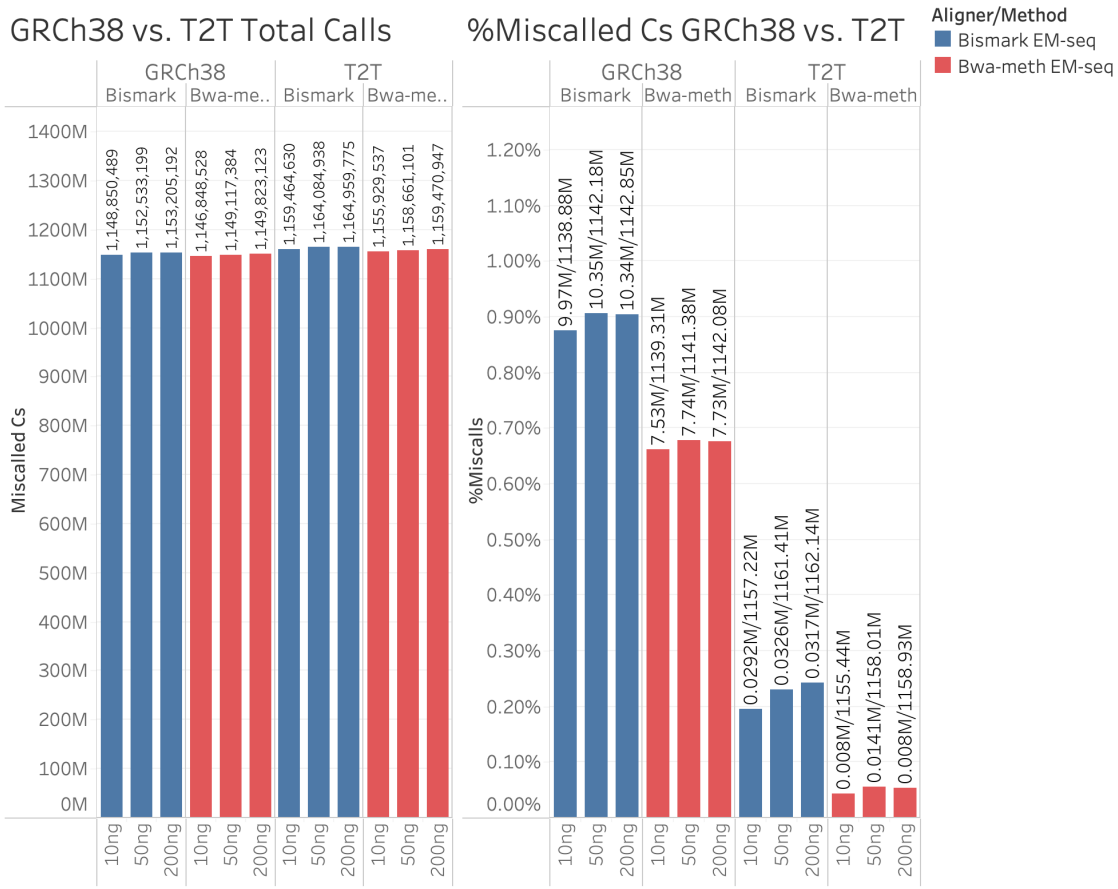

Figure 12: Left: Total call counts before filtering in the 10, 50, and 200ng EM-seq samples, aligned with bwa-meth to GRCh38 and T2T. The vertical axis is the number of calls in the sample. Right: Percents of miscalled Cs in the same samples, where a "miscall" is a methylation call filtered out entirely by per-read filtering. The vertical axis is the percent of miscalled Cs in the sample. In both figures, the colors define which sequencing method and aligner were used, as detailed in the legend.

Sequencing Method	Sample Mass	Read Count
WGBS	10ng	1,268,442,060
WGBS	50ng	1,269,601,144
WGBS	200ng	1,269,737,220
EM-seq	10ng	1,298,351,192
EM-seq	50ng	1,296,831,526
EM-seq	200ng	1,296,793,470

Table 1: Sequencing read counts broken down by sequencing method and sample mass. Read counts are reported as counts of individual reads, not pairs.

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 14 and 15. Tools used include bedtools [26], samtools [19], GNU awk, MethylDackel, bigWigToBedGraph [13], BEDOPS [22], 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 v0.6.0. MethylDackel uses a default value of 10 as a minimum MapQ which should include mostly single locus reads, however inaccurate MapQs can 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. MethylDackel's trimming settings were set to trim off the first 1 bp of read 1 and the last 2 bp of read 2.

The ClinVar VCF and GENCODE GFF were combined using bedtools and

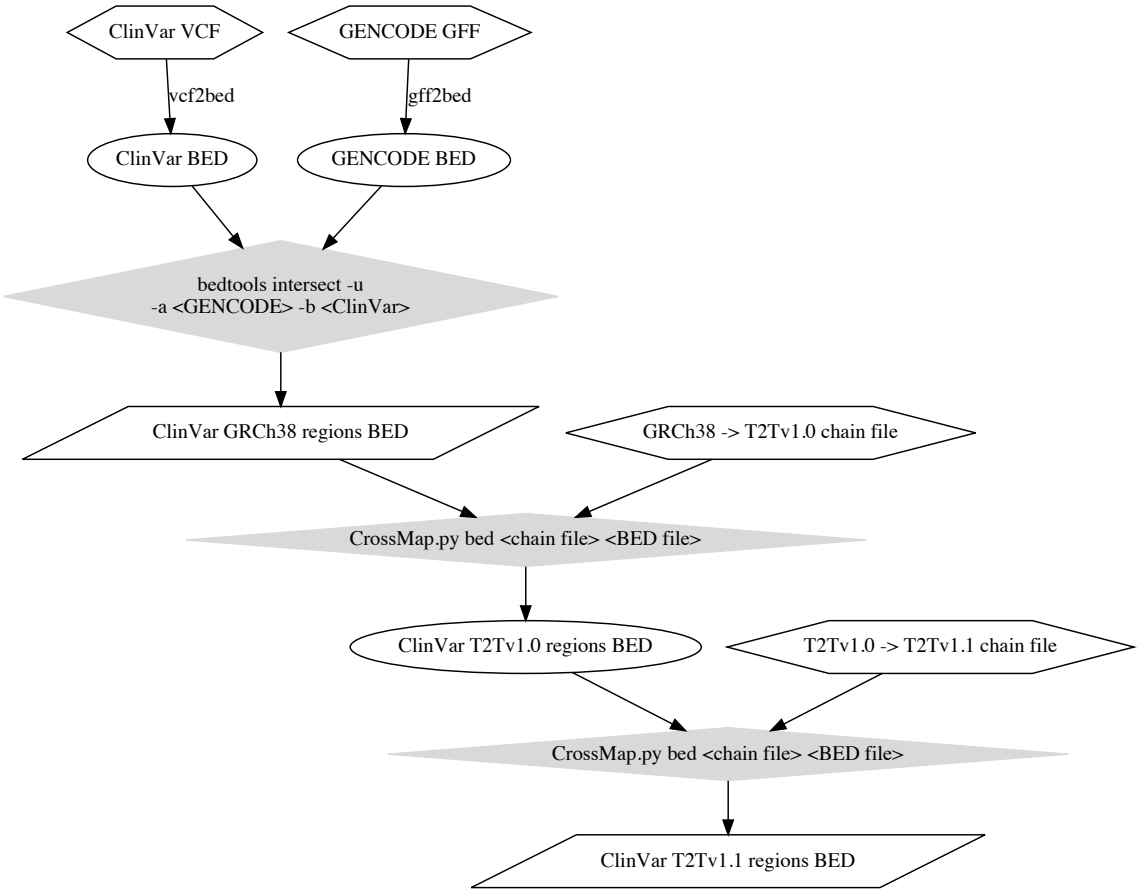

Figure 13: Detailed description of the procedure used to annotate the T2Tv1.1 reference using the GRCh38 ClinVar and GENCODE data. Hexagons are input files. Ovals are intermediate files. Gray rhombuses are processing steps with multiple arguments. Parallelograms are output files. Labels on edges are processing steps with one input and one output.

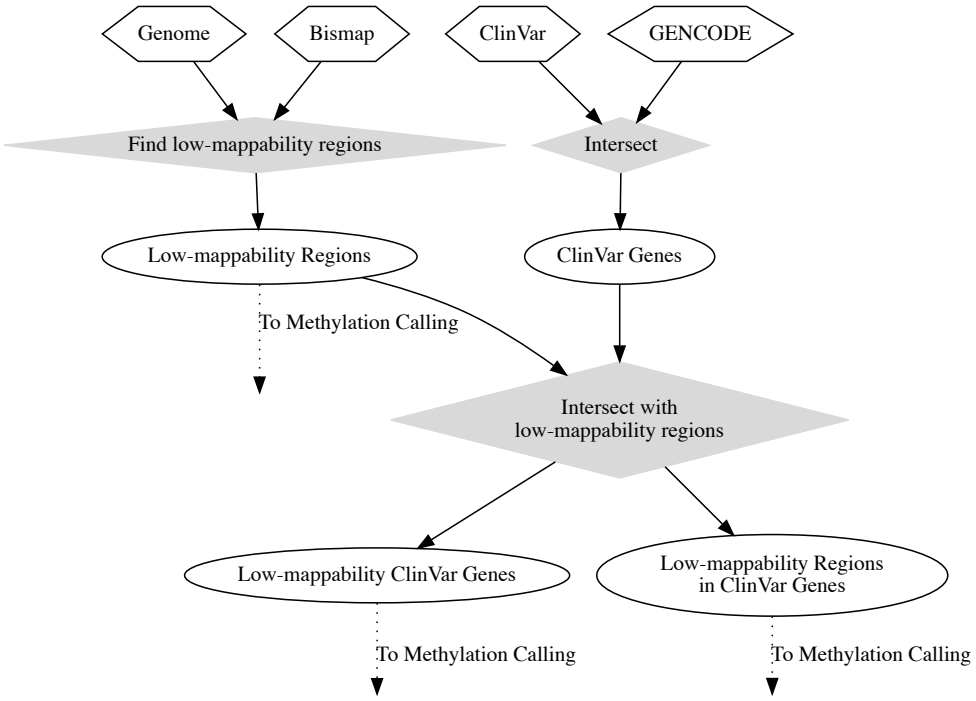

Figure 14: Detailed description of preprocessing steps used to identify biologically relevant low-mappability regions.. Ovals are intermediate files. Gray rhombuses are processing steps with multiple arguments. Dotted lines indicate inputs that go to the methylation calling steps in Figure 15.

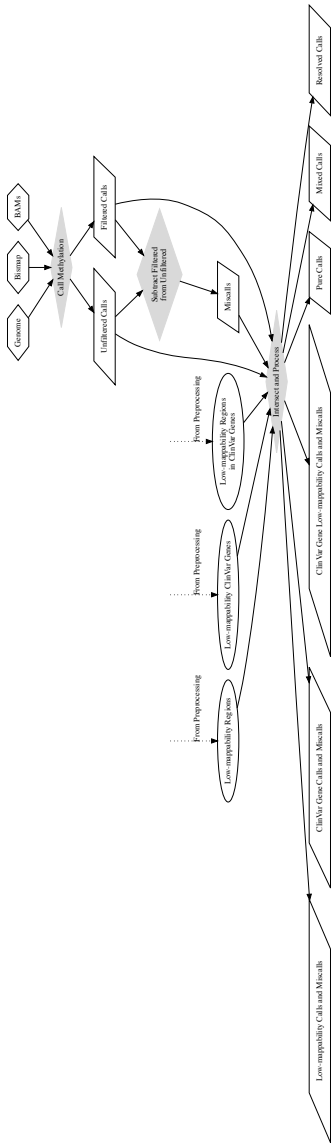

Figure 15: Detailed description of steps for calling methylation and intersecting the resulting calls with the low mappability regions determined in the preprocessing steps. Ovals are intermediate files. Gray rhombuses are processing steps with multiple arguments. Parallelograms are output files. Dotted lines indicate inputs that come from the preprocessing steps in Figure 14. Parallelograms are output files.

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". Each region in the ClinVar regions BED was expanded 2kb upstream using bedtools slop to capture methylation upstream of the gene. The Bismap bedGraph (which 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 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 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 methylation calls in low mappability ClinVar regions (this will be referred to as the "low mappability ClinVar calls file"). A low-mappability methylation call, as used here, is defined as a methylation call that is in a region with Bismap mappability less than 0.01, as defined by the position of the C alone (positions of anchoring reads are not considered when determining this).

The low mappability ClinVar calls file was intersected with the file of ClinVar regions with low minimum mappability to produce a file of ClinVar regions with low mappability calls. The -wa option for bedtools intersect was used here, which writes a copy of the ClinVar region to the output file for each low mappability call in the region, in order that this file would contain multiple copies of each region, one per low mappability call in the region. These duplicates were then used to count low mappability calls 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 low mappability calls and how many low mappability calls are in each region. A similar intersection was performed for all ClinVar regions with low minimum mappability (i.e. ClinVar regions that could potentially be affected) to produce a file of all calls (regardless of mappability of the C) in such regions, and a list of such regions with a count of calls per region.

Since this analysis was run for three different input masses and two sequencing protocols, the low mappability calls files were also processed through a custom Python script (see supplemental materials) to add a field specifying which input mass, sequencing protocol, and MethylDackel filtering setting were used. As the input mass would have been difficult to parse out of the BAM name due to inconsistent file name formatting, a CSV was created mapping BAM file

names to input masses and given as input to this step. The same field was added to the lists of all ClinVar regions with low mappability calls and counts described previously. We also produced similarly-annotated files for all methylation calls (regardless of mappability), all methylation calls in GENCODE genes (with low minimum mappability) which contained ClinVar variants, all methylation calls in low-mappability regions that were also in GENCODE genes which contained ClinVar variants, counts of calls in in GENCODE genes which contained ClinVar variants regardless of mappability, methylation calls that are either 0% or 100% (regardless of mappability), and methylation calls which are neither 0% nor 100% (also regardless of mappability)..

To assess miscalls (i.e. which Cs are filtered out by per-read filtering), as well as underfiltering/overfiltering with respect to per-C filtering (discussed above), and resolution of mixed methylation calls to 0% or 100%, pairs of these files were compared using bedtools to produce files showing the difference between filtered and unfiltered data. 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 using default settings and deduplicated using deduplicate_bismark according to the documented protocol [14]. The name of the aligner was added to the field specifying the input mass, protocol, and MethylDackel filtering settings present in all output files containing methylation calls or counts of such calls in ClinVar regions (using the same custom Python script).

To compare the need for and the effect of mappability filtering on the more complete T2T v1.1 reference genome [23] (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 datasets, 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 CrossMap [32]. This lift-over was done in two steps with separate chain files. First, the GRCh38 regions were lifted over using CrossMap to T2Tv1.0 using a chain file published on GitHub by Nico Alavi [1]. Then, the lifted-over T2Tv1.0 annotations were lifted over again to T2Tv1.1 using a chain file from the T2T Consortium [20]. Specifically, the ClinVar regions BED, with the ClinVar and GENCODE data already combined, was lifted over. This converted file was used in place of the ClinVar regions BED for analysis with the T2T reference. As there was no 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.In total, this analysis was run on all combinations of sequencing method (WGBS or EM-seq), aligner (Bismark or bwa-meth), input mass (10, 50, or 200 ng), filtering (read-based filtering or no read-based filtering), and reference genome (GRCh38 or T2T), using a Nextflow[5] pipeline (see supplemental materials) which managed the execution of all the analysis tools needed.

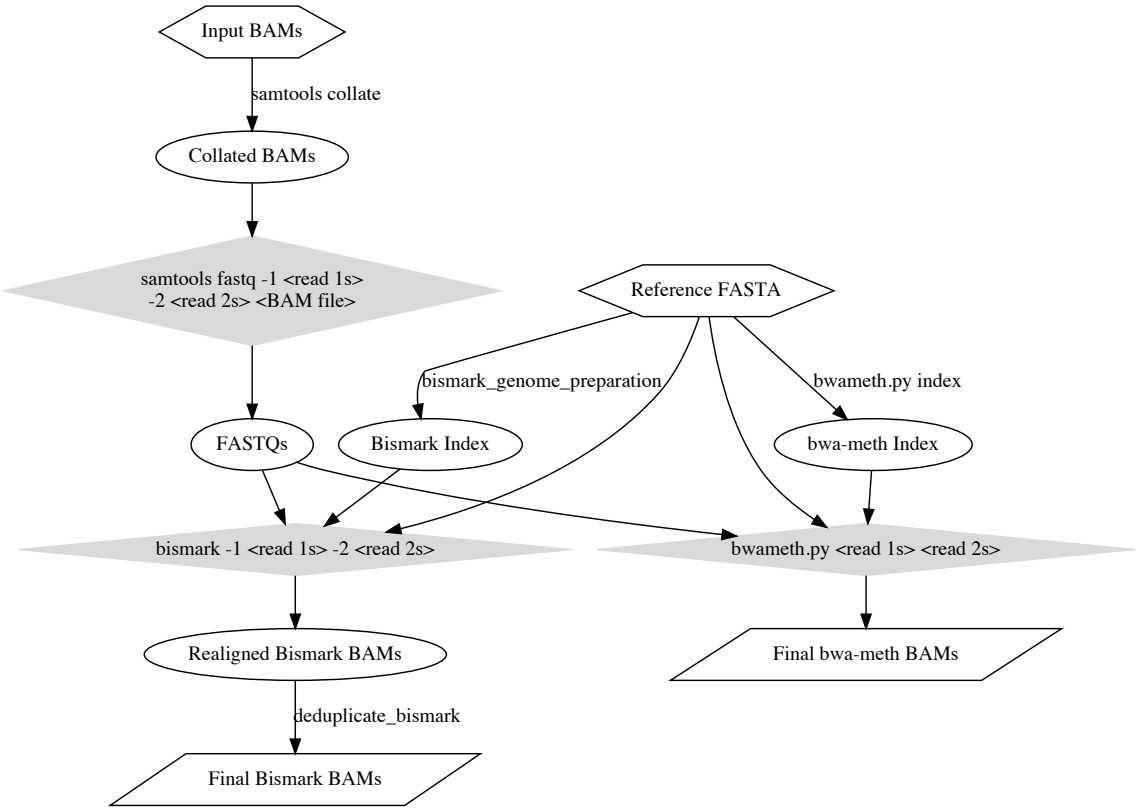

Figure 16: Detailed description of BAM realignment steps. Hexagons are input files. Ovals are intermediate files. Gray rhombuses are processing steps with multiple arguments. Parallelograms are output files. Labels on edges are processing steps with one input and one output.

Discussion

Reads placed with falsely high confidence have cascading detrimental effects on methylation calling, differential methylation assessment, and assessment of phenotypes associated with methylation status. To remove the spurious, unsupported methylation calls that result from such reads, we suggest that reads with both mates in low mappability regions (as determined by Bismap) should be excluded from methylation calling. Additionally, because of the more accurate MapQ values, decreased run time, and more flexibility to separate methylation calling from alignment, we also recommend the use of bwa-meth for alignment and MethylDackel with MapQ > 10 for methylation calling.

Supplemental Materials

The Nextflow scripts used to analyze this data can be found at `https://github.com/nebiolabs/low_bismap_methyl_calls` . Individual tools used in the analysis include sambamba[30], vcf2bed, gff2bed, bigWigToBedGraph, MethylDackel, awk, python, GNU sort, GNU head, bedtools, and BEDOPS. Versions of all tools used are specified in the nextflow 21.04.0 scripts using conda[21] dependency resolution.

The custom Python script that adds the input mass, sequencing protocol, MethylDackel filtering setting, and aligner name can be found at `https://github.com/nebiolabs/low_bismap_methyl_calls`.

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

The chromosome ideogram figure was generated using karyoploteR[7].

References

[1] Nico Alavi. *burgshrimps/liftover_ T2T*. original-date: 2020-11-03T21:50:35Z. May 2021. URL: `https://github.com/burgshrimps/liftover_T2T` (visited on 07/07/2021).

[2] Achim Breiling and Frank Lyko. "Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond". In: *Epigenetics & Chromatin* 8 (2015), p. 24. ISSN: 1756-8935.

[3] Gary G. Chen et al. "Medium throughput bisulfite sequencing for accurate detection of 5-methylcytosine and 5-hydroxymethylcytosine". In: *BMC Genomics* 18 (Jan. 2017). 28100169[pmid], p. 96. ISSN: 1471-2164.

[4] Deanna M. Church et al. “Modernizing reference genome assemblies”. eng. In: *PLoS biology* 9.7 (July 2011), e1001091. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001091.

[5] Paolo Di Tommaso et al. “Nextflow enables reproducible computational workflows”. en. In: *Nature Biotechnology* 35.4 (Apr. 2017), pp. 316–319. ISSN: 1546-1696. DOI: 10.1038/nbt.3820. URL: https://www.nature.com/articles/nbt.3820 (visited on 07/12/2021).

[6] Francine E. Garrett-Bakelman et al. “Enhanced Reduced Representation Bisulfite Sequencing for Assessment of DNA Methylation at Base Pair Resolution”. In: *J Vis Exp* 96 (Feb. 2015). 25742437[pmid], p. 52246. ISSN: 1940-087X.

[7] Bernat Gel and Eduard Serra. “karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data”. In: *Bioinformatics* 33.19 (Oct. 2017), pp. 3088–3090. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx346. URL: https://doi.org/10.1093/bioinformatics/btx346 (visited on 08/12/2021).

[8] The SAM/BAM Format Specification Working Group. *Sequence Alignment/Map Format Specification*. English. https://samtools.github.io/hts-specs/SAMv1.pdf. May 2018. (Visited on 08/09/2018).

[9] Jennifer Harrow et al. “GENCODE: The reference human genome annotation for The ENCODE Project”. In: *Genome Res* 22.9 (Sept. 2012). 22955987[pmid], pp. 1760–1774. ISSN: 1088-9051.

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

[11] Johnurbangenome. *Biofinysics: How does bowtie2 assign MAPQ scores?* May 2014. URL: http://biofinysics.blogspot.com/2014/05/how-does-bowtie2-assign-mapq-scores.html (visited on 06/18/2021).

[12] Mehran Karimzadeh et al. “Umap and Bismap: quantifying genome and methylome mappability”. In: *Nucleic Acids Research* (2018).

[13] W. J. Kent et al. “BigWig and BigBed: enabling browsing of large distributed datasets”. In: *Bioinformatics* 26.17 (2010). http://genome.ucsc.edu/, pp. 2204–2207. eprint: /oup/backfile/content_public/journal/bioinformatics/26/17/10.1093_bioinformatics_btq351/2/btq351.pdf.

[14] Felix Krueger. *FelixKrueger/Bismark*. original-date: 2015-11-07T18:14:13Z. July 2021. URL: https://github.com/FelixKrueger/Bismark (visited on 07/13/2021).

[15] Felix Krueger and Simon R. Andrews. “Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications”. In: *Bioinformatics* 27.11 (2011), pp. 1571–1572. eprint: /oup/backfile/content_public/journal/bioinformatics/27/11/10.1093_bioinformatics_btr167/1/btr167.pdf.

[16] Tsong-teh Kuo, Tan-chi Huang, and Mei-hui Teng. "5-Methylcytosine replacing cytosine in the deoxyribonucleic acid of a bacteriophage for Xanthomonas oryzae". en. In: *Journal of Molecular Biology* 34.2 (June 1968), pp. 373–375. ISSN: 0022-2836. DOI: 10.1016/0022-2836(68)90263-5. URL: https://www.sciencedirect.com/science/article/pii/0022283668902635 (visited on 07/30/2021).

[17] Melissa J Landrum et al. "ClinVar: improving access to variant interpretations and supporting evidence". In: *Nucleic Acids Research* 46.D1 (2018), pp. D1062–D1067. eprint: /oup/backfile/content_public/journal/nar/46/d1/10.1093_nar_gkx1153/2/gkx1153.pdf.

[18] Yan-Jiun Lee and Peter R. Weigele. "Detection of Modified Bases in Bacteriophage Genomic DNA". en. In: *DNA Modifications: Methods and Protocols*. Ed. by Alexey Ruzov and Martin Gering. Methods in Molecular Biology. New York, NY: Springer US, 2021, pp. 53–66. ISBN: 978-1-07-160876-0. DOI: 10.1007/978-1-0716-0876-0_5. URL: https://doi.org/10.1007/978-1-0716-0876-0_5 (visited on 07/13/2021).

[19] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. eprint: /oup/backfile/content_public/journal/bioinformatics/25/16/10.1093/bioinformatics/btp352/2/btp352.pdf.

[20] *marbl/CHM13*. original-date: 2019-02-28T16:00:16Z. July 2021. URL: https://github.com/marbl/CHM13 (visited on 07/07/2021).

[21] *Miniconda*. Apr. 2021. URL: https://docs.conda.io/en/latest/.

[22] Shane Neph et al. "BEDOPS: high-performance genomic feature operations". In: *Bioinformatics* 28.14 (2012), pp. 1919–1920. eprint: /oup/backfile/content_public/journal/bioinformatics/28/14/10.1093_bioinformatics_bts277/2/bts277.pdf.

[23] Sergey Nurk et al. "The complete sequence of a human genome". en. In: *bioRxiv* (May 2021). Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2021.05.26.445798. DOI: 10.1101/2021.05.26.445798. URL: https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1 (visited on 07/07/2021).

[24] B. S. Pedersen et al. "Fast and accurate alignment of long bisulfite-seq reads". In: *ArXiv e-prints* (Jan. 2014). arXiv: 1401.1129 [q-bio.GN].

[25] *QC Fail Sequencing Â» MAPQ values are really useful but their implementation is a mess*. en. URL: https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/ (visited on 06/18/2021).

[26] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842. eprint: /oup/backfile/content_public/journal/bioinformatics/26/6/10.1093_bioinformatics_btq033/3/btq033.pdf.

[27] James T. Robinson et al. “Integrative genomics viewer”. en. In: *Nature Biotechnology* 29.1 (Jan. 2011), pp. 24–26. ISSN: 1546-1696. DOI: 10.1038/nbt.1754. URL: https://www.nature.com/articles/nbt.1754 (visited on 08/09/2021).

[28] Devon Ryan. *MethylDackel*. English. https://github.com/dpryan79/MethylDackel. (Visited on 08/09/2018).

[29] Dirk Schübeler. “Function and information content of DNA methylation”. In: *Nature* 517 (Jan. 2015), 321 EP -.

[30] Artem Tarasov et al. “Sambamba: fast processing of NGS alignment formats”. In: *Bioinformatics* 31.12 (2015), pp. 2032–2034. DOI: 10.1093/bioinformatics/btv098. URL: +%20http://dx.doi.org/10.1093/bioinformatics/btv098.

[31] Romualdas Vaisvila et al. “Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA”. en. In: *Genome Research* (June 2021). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.266551.120. URL: https://genome.cshlp.org/content/early/2021/06/17/gr.266551.120 (visited on 07/13/2021).

[32] Hao Zhao et al. “CrossMap: a versatile tool for coordinate conversion between genome assemblies”. eng. In: *Bioinformatics (Oxford, England)* 30.7 (Apr. 2014). Edition: 2013/12/18 Publisher: Oxford University Press, pp. 1006–1007. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt730. URL: https://pubmed.ncbi.nlm.nih.gov/24351709.