


# Accelerating the alignment processing speed of the comprehensive end-to-end whole-genome bisulfite sequencing pipeline, wg-blimp

Jake D. Lehle <sup>1,\*</sup> and John R. McCarrey<sup>1</sup>

<sup>1</sup>Department of Neuroscience, Developmental and Regenerative Biology, The University of Texas at San Antonio, San Antonio, TX 78249, USA

\*Correspondence address. Department of Neurosciences, Developmental and Regenerative Biology, The University of Texas at San Antonio, 1 UTSA Circle, San Antonio, TX 78249, USA. Tel: +1 (512)-992-8144; E-mail: [jake.lehle@utsa.edu](mailto:jake.lehle@utsa.edu)

## Abstract

Analyzing whole-genome bisulfite and related sequencing datasets is a time-intensive process due to the complexity and size of the input raw sequencing files and lengthy read alignment step requiring correction for conversion of all unmethylated Cs to Ts genome-wide. The objective of this study was to modify the read alignment algorithm associated with the whole-genome bisulfite sequencing methylation analysis pipeline (wg-blimp) to shorten the time required to complete this phase while retaining overall read alignment accuracy. Here, we report an update to the recently published pipeline wg-blimp achieved by replacing the use of the bwa-meth aligner with the faster gemBS aligner. This improvement to the wg-blimp pipeline has led to a more than  $\times 7$  acceleration in the processing speed of samples when scaled to larger publicly available FASTQ datasets containing 80–160 million reads while maintaining nearly identical accuracy of properly mapped reads when compared with data from the previous pipeline. The modifications to the wg-blimp pipeline reported here merge the speed and accuracy of the gemBS aligner with the comprehensive analysis and data visualization assets of the wg-blimp pipeline to provide a significantly accelerated workflow that can produce high-quality data much more rapidly without compromising read accuracy at the expense of increasing RAM requirements up to 48 GB.

**Keywords:** whole-genome bisulfite sequencing; analysis pipeline; epigenetics; DNA methylation

## Introduction

DNA methylation at CpG dinucleotides is one of the most commonly studied parameters within the epigenome because it is directly assessable and is often reflective of the overall structure of chromatin, which, in turn, contributes to regulation of gene expression at the transcriptional level [1]. While there is a myriad of techniques for analysis of DNA methylation, a number of those used in the past (e.g. reduced-representation bisulfite sequencing [2], methylated DNA immunoprecipitation sequencing [3]) have employed enrichment of regions with higher frequencies of CpG dinucleotides to limit the portion of the genome to be sequenced as a means to decrease the cost and computational resources required to process and analyze the resulting data. However, these techniques provide only a partial view of the epigenome, typically focused primarily on the impact of DNA methylation on chromatin structure in promoters and exons where CpG dinucleotides are often most abundant [4]. This limits the potential of these techniques to profile DNA methylation in other regions of the genome which also contribute to regulation of gene expression, such as enhancers or regions associated with the boundaries of topologically associated domains [5].

Whole-genome approaches, such as whole-genome bisulfite sequencing (WGBS) and more recently, enzymatic methyl-seq (EM-seq), yield informative results for the entire genome, and, as such, have become the gold standard for global analysis of DNA

methylation with single-CpG resolution [6]. Thus, as sequencing costs have decreased [7], an increasing number of investigators are opting to utilize this more comprehensive, genome-wide assessment of DNA methylation which yields large and robust datasets [8–10]. However, this more comprehensive assessment of the epigenome mandates a corresponding increase in the extent of computational analysis needed to interpret the resulting larger datasets.

Recently, a novel snakemake [11] workflow termed whole-genome bisulfite sequencing methylation analysis pipeline (wg-blimp) was described as an “end-to-end” pipeline for processing WGBS data by integrating established algorithms for alignment, quality control (QC), methylation calling, detection of differentially methylated cytosines (DMCs), differentially methylated regions (DMRs), and methylation segmentation for profiling of DNA methylation states at regulatory elements [12]. This novel wg-blimp pipeline has been added to a short but growing list [13–17] of similar pipelines which include msPIPE [18], the ENCODE WGBS pipeline [19], and Nextflow methylseq [20]. The wg-blimp pipeline is simple to install on either a personal computer or in a research high-computing cluster, often requiring only an input reference, gene annotation, and FASTQ read files to fully process WGBS data. However, due to the nature and large file sizes of WGBS data, executing the wg-blimp pipeline in its previous form often required extended computing time emanating from

Received: April 18, 2023. Revised: June 12, 2023. Accepted: June 12, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

mapping bisulfite-treated sequences to a reference genome. This difficulty is due to the conversion of unmethylated Cs to Us in the original DNA strand following bisulfite treatment. During PCR amplification, these Us are replaced with Ts, ultimately resulting in the conversion of C–G base pairs into T–A base pairs. Because most Cs in the genome exist in non-CpG contexts and are thus normally unmethylated, the bisulfite treatment causes a substantial increase in the proportion of T–A base pairs and a concomitant decrease in the proportion of G–C base pairs in the amplified copies of the initially treated DNA strands. This renders mapping of bisulfite-converted reads using a conventional read mapper inadequate, because a large percentage of the converted bases will be called as mismatches relative to the untreated reference sequence.

To overcome this limitation, improved “3-letter” aligners, such as *bwa-meth* [21] and *gemBS* [22], designed specifically for mapping bisulfite-converted reads, perform a two-stage mapping process. Cs on read 1 are fully converted to Ts, while Gs on read 2 are fully converted to As. The reads are then aligned to either of two reference genomes, where either all the Cs have been converted to Ts or all Gs have been converted to As. After mapping to the converted reference genomes, the read sequences are then restored to the original sequence, revealing methylated Cs which can be identified in further downstream processing. Due to this extensive processing step required for alignment of all converted reads to multiple indexed genomes, followed by conversion back to the starting read sequence, the alignment step imposes a very time-consuming computational toll on data processing.

While both *bwa-meth* and *gemBS* follow the same “3-letter” alignment mapping concept, there are significant differences in their implementation which translate to large differences in their overall speed due to differences in the underlying alignment software packages from which these specialized methylation aligners were generated. The *bwa-meth* methylated DNA aligner has a foundation built on the improved BWA-MEM alignment software which follows the seed-and-extend paradigm to find initial seed alignment with super-maximal exact matches (SMEMs) using an improvement of the Burrows–Wheeler transform algorithm [21, 23]. BWA-MEM additionally re-seeds SMEMs greater than the default of 28 bp to find the longest exact match in the middle of the seed that occurs at least once in the bisulfite-converted reference genome, to reduce potential miss-mapping due to missing seed alignments. BWA-MEM also filters out unneeded seeds by grouping closely located seeds which it terms ‘chains’, thereby removing shorter chains contained within longer chains (which are at least 50% and 38 bp shorter than the longer chain) [23]. The seeds remaining in these longer chains are then ranked by the length of the chain to which the seed belongs, and then by the length of the seed itself. Seeds that are already contained in a previously identified alignment are dropped, while seeds that potentially lead to a new alignment are extended with a banded affine-gap-penalty dynamic program [23]. While these strategies have increased the potential size of the read that can be aligned using the BWA-MEM software from 70 bp to a few megabases, the heavy reliance on gapped sequenced alignment by BWA-MEM comes with the drawback that non-unique matches have a higher likelihood of aligning to multiple places in the genome introducing a higher potential for false-positive read alignments.

While the aligner that the *gemBS* software is built on, GEM3, allows for mapping lengths of only up to 1 kb, this length is sufficient to scale to large sequencing analyses while GEM3 prioritization of exact over gapped read alignments maintains equal if not superior read mapping accuracy when compared with BWA-MEM

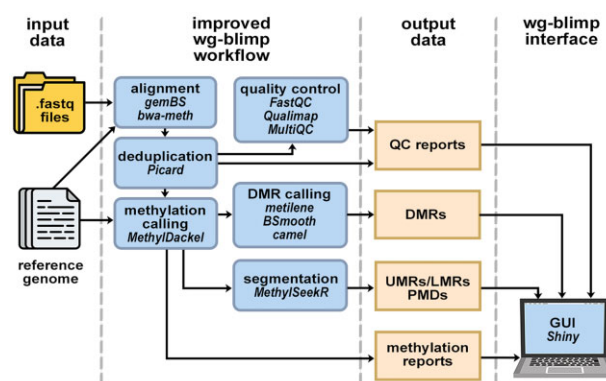
[22]. This superiority largely comes from *gemBS* performing the conversion of read steps before and after mapping “on the fly” [22] for each read pair, thereby avoiding the generation of intermediate files and greatly increasing the efficiency of the mapping process. In addition, GEM3 filters and sorts mapped seeds into groups referred to as “strata” which facilitate complete searches of indexed references to find all possible matches to the reference genome, improving both the speed and the accuracy over BWA-MEM and other heuristic mapping algorithms but coming at the cost of filtering away a larger number of read strings that cannot be grouped into seed groups compared with BWA-MEM [24].

Searching through such a large index file does expose one additional limitation of *gemBS*, which is that it requires 48 GB of RAM compared with only 8–16 GB required by *bwa-meth*. However, this limitation is normally insignificant given that most midrange or higher computers are equipped with more than sufficient RAM to meet this need [22]. We sought to leverage these differences to improve the speed of read alignment in the *wg-blimp* pipeline. We were able to modify the *wg-blimp* pipeline by replacing the *bwa-meth* alignment software with *gemBS*. This single modification allowed us to increase the overall speed of the *wg-blimp* pipeline by more than  $\times 7$ , all without sacrificing alignment accuracy.

## Materials and methods

### Wg-blimp workflow steps

The aim of the *wg-blimp* pipeline is to provide a straightforward and streamlined framework for processing WGBS or similar data that will allow the user to perform genome-wide DNA methylation analysis by bundling a suite of preexisting software packages. The result is an end-to-end inclusive pipeline that is useable by biologists of all skill levels. Users without an extensive computer science background will appreciate the comprehensive walkthroughs available as well as the limited time and files required to install and begin running the pipeline, while more experienced users will appreciate the tuning that can be configured among multiple steps to customize analysis without extensive effort. Figure 1 shows an overview of the analysis steps included in the pipeline.



**Figure 1:** Improved *wg-blimp* workflow overview. FASTQ files and the reference genome file provided by the user now have the option to be aligned by the newly added *gemBS* aligner. Output BAM files are then processed through the remainder of the *wg-blimp* pipeline and results can be viewed using the web browser interface.

### Step 1: input data

To start the workflow, users processing human samples need only provide the FASTQ file and a reference genome as input. This can be done directly from the command line when running the *wg-blimp* binary following installation. Alternatively, paths to the FASTQ and reference genome files can be added to a workflow config file that can be customized for each application or species being analyzed.

### Step 2: *wg-blimp* workflow

Alignment of FASTQ files can now be toggled between the newly added *gemBS* or previous *bwa-meth* aligner, either in the command line or master config file when running the pipeline. As both *gemBS* and *bwa-meth* have internal usage of soft-clipping of reads to mask non-masking read sequences, alignment pre-trimming is still omitted from the *wg-blimp* pipeline and is an optional pre-processing step to help improve alignment of reads. Following alignment, duplicates produced from PCR amplification of reads during library preparation are removed using the Picard toolkit [25].

The pipeline contains a number of programs for QC of the output BAM files from alignment. Read quality scores are determined with FastQC [26] and Qualimap [27] is used to determine the overall and per-chromosome read coverage. In addition, Qualimap also calculates GC content, duplication rate, and clipping profiles. Plots for visualization of all QC metrics are aggregated into a single user-friendly HTML report using MultiQC [28] discussed in more depth in Step 3.

Methylation calling is performed by MethylDackel [29] which creates a methylation report used to compute global methylation statistics. Computing the C > T conversion rate can be enabled by computing per-chromosome methylation, as unmethylated lambda DNA is commonly added to genomic DNA prior to bisulfite treatment as an internal control. The pipeline contains multiple algorithms for DMC/DMR calling including: metilene [30], BSmooth [31], and camel [32] which all can be used in parallel while running the pipeline and were intended to help strengthen the biological significance of overlapping DMCS/DMRs which were identified by different approaches. The parameters that are used to define DMRs can all be tuned by the user and include limits to the minimum number of CpG sites that must be differentially methylated within the region, the minimum coverage size for each DMR, and minimum average difference in methylation between two groups. In addition, metilene also includes *q*-values computed from the Mann–Whitney *U* test that can be used to filter DMRs based on significance during data analysis. Individual DMCS can be identified by setting the minimum number of differentially methylated CpGs in the region to one.

*Wg-blimp* can identify potential regulatory regions that may be bound by transcription factors leading to a reduction in DNA methylation through the use of MethylSeekR [33] which can detect both unmethylated regions (UMRs) and low-methylated regions (LMRs) based off a user-defined false-discovery rate (FDR) and methylation cutoff. MethylSeekR also detects regions that show a high level of disorder in DNA methylation which it defines as partially methylated domains (PMDs). However, these PMDs can influence and limit the detection of LMRs and UMRs. So, to remove this influence, *wg-blimp* performs the MethylSeekR workflow with and without PMD identification and users can ultimately decide whether or not to include PMDs when analyzing UMRs and LMRs.

Finally, the resulting DMRs, UMRs, LMRs, and PMDs are all annotated for overlap with genes, promoters, CpG islands, and repetitive elements from the respective Ensemble [34], UCSC [35], and RepeatMasker [36] databases. A final QC check is performed by computing the average coverage per DMR with *mosdepth* [37], so that regions of low coverage can be filtered out and removed from further analysis.

### Step 3: output visualization interface

Upon completion of the pipeline, users may visualize all of the data from their WGBS experiment in the *wg-blimp* interactive results web browser that is built using the R Shiny local browser hosting framework [38]. The straightforward layout of the *wg-blimp* interactive results browser is user-friendly, allowing users to quickly find information about the QC reports, pipeline settings, DMRs, and segmentation broken down into separate tabs. From the browser, users can dynamically adjust the data to filter both the DMRs and UMRs/LMRs results tables as well as choosing whether or not to include PMDs as described above. In addition, all of the results tables and publication quality figures produced within the browser are easily exported, streamlining and simplifying data processing.

### Runtime benchmarking, stress testing, and accuracy testing of improved *wg-blimp* v0.10.0

We performed a series of side-by-side runtime benchmarking, stress testing, and accuracy testing in order to determine if adding the *gemBS* aligner to the *wg-blimp* pipeline would accelerate the runtime while maintaining the accuracy of read alignment. We first compared overall runtime on WGBS sample datasets provided to test the *wg-blimp* pipeline installation, which included isogenic human blood and sperm WGBS files (each generated from pools of DNA from six men) with nearly 1 million (M) reads each, all restricted to chr22 [12, 39]. We then tested the runtime of the two pipeline versions on larger publicly available mouse WGBS data from CD19+ B cells [40] and spermatocytes [41] with files ranging in size from 80 to 160 M reads to model file sizes that are more typical to analysis of current WGBS or similar datasets.

We next tested the extent to which any increase in alignment speed afforded by the use of *gemBS* might be accompanied by reduced alignment accuracy. First, we stress-tested each aligner by comparing the mapping percentages of three groups of simulated paired-end converted sequence reads produced by the *mason2* application [42] (Supplementary Section S1) with the number of reads in each group increasing by a log scale from 1 to 100 M reads. We defined accuracy as the percentages of both mapped and unique or ‘properly paired’ reads indicated in the BAM file output for each aligner using the *flagstat* function from SAMtools [43]. In addition, we also took into consideration other metrics, including the number of reads that had their mate read mapped to a different chromosome or had a low overall mapping QC score. However, there is the potential for either of the aligners to output false-positive read mapping that could bias the perceived accuracy of one aligner over the other. To account for this potential bias, we also tested the ability of each aligner to accurately map reads that could be used to identify a previously known DMR as an additional biologically relevant test of accuracy.

To further assess the accuracy of these aligners in a biological context, we obtained publicly available mouse WGBS data from CD19+ B cells [40] and spermatocytes [41] and used each aligner to analyze DNA methylation patterns in each dataset focusing specifically on methylation of the *Pgk2* gene promoter. *Pgk2* is an



intronless gene that arose via retrotransposition of the *Pgk1* gene and is required during normal spermatogenesis [44, 45]. Our lab has previously shown that in mice, the upstream half of the *Pgk2* gene becomes demethylated in prospermatogonia and spermatogonia, prior to the activation of transcription of the *Pgk2* gene in spermatocytes [46, 47]. Thus, we tested the accuracy with which pipelines utilizing each aligner software were able to correctly identify this known DMR.

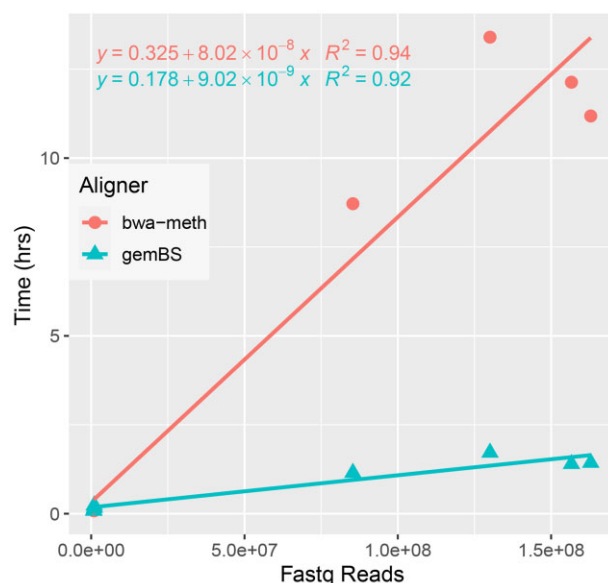
While this comparison allowed us to test for the identification of a single expected DMR as a biological measure of accuracy, we also sought to determine how changing the alignment algorithm could alter the identification of methylated regions throughout the remainder of the epigenome. To determine whether any differences in the alignment of reads by either aligner could translate into differences in the clustering and identification of methylation regions, we also compared the output of all other DMRs, UMRs, LMRs, and PMDs produced by each aligner for the mouse CD19+ B cells and spermatocytes. The statistical significance of any differences between each group was tested with Pearson's Chi-squared and Kruskal-Wallis rank sum tests. This comparison gave us an epigenome-wide view of how altering the alignment algorithm can impact global methylated region identification which should be considered by users when analyzing DNA methylation data.

Finally, to illustrate the ability for the pipeline to handle not only WGBS data but also increasingly utilized EM-seq datasets, we tested our improved pipeline on WGBS and EM-seq data from human genomic DNA and cell-free DNA found in blood samples collected from cancer patients (Supplementary Section S2) [48]. EM-seq utilizes enzymatic conversion based on TET2/APOBEC2 which produces the same modifications as chemical bisulfite-based conversion methods and has been gaining popularity because enzymatic conversion of cytosines avoids the fragmentation of DNA typically generated by bisulfite conversion of DNA. The latter occurs under acidic conditions (pH 5) and at high temperatures (90°C) which are not required for the EM-seq method, thus allowing this enzymatic method to generate libraries with longer DNA inserts [49].

All analyses were executed on a server equipped with one Intel Cascade Lake CPU with 80 physical cores and 160 hyper-threads or virtual cores, 394 GB of memory, and a CentOS 7 operating system. When running the pipeline, we limited the core utilization to only 8 of the 80 available cores, which we found was nearing a plateau in multithreading efficiency for alignment speed (Supplementary Section S3). Links to all files and code we used to run the analysis can be found in the Data and Code Accessibility and Supplementary Data sections.

## Results

Benchmarking in previous studies has shown that gemBS is a superior alignment software with respect to overall mapping processing time, because it can scale for use with larger datasets more effectively than bwa-meth [22, 50, 51]. We modified the wg-blimp pipeline to replace the bwa-meth aligner with gemBS, and then tested our prediction that this change would lead to a decrease in the time required for the alignment step in the pipeline (Fig. 1). When comparing the runtime of the smaller practice files with 1 M reads provided with the wg-blimp pipeline, we found the average bwa-meth alignment time was only slightly shorter than the gemBS alignment time. However, when running the pipeline with FASTQ files containing 80–160 M reads each, we observed a large difference in the time required to align each file,

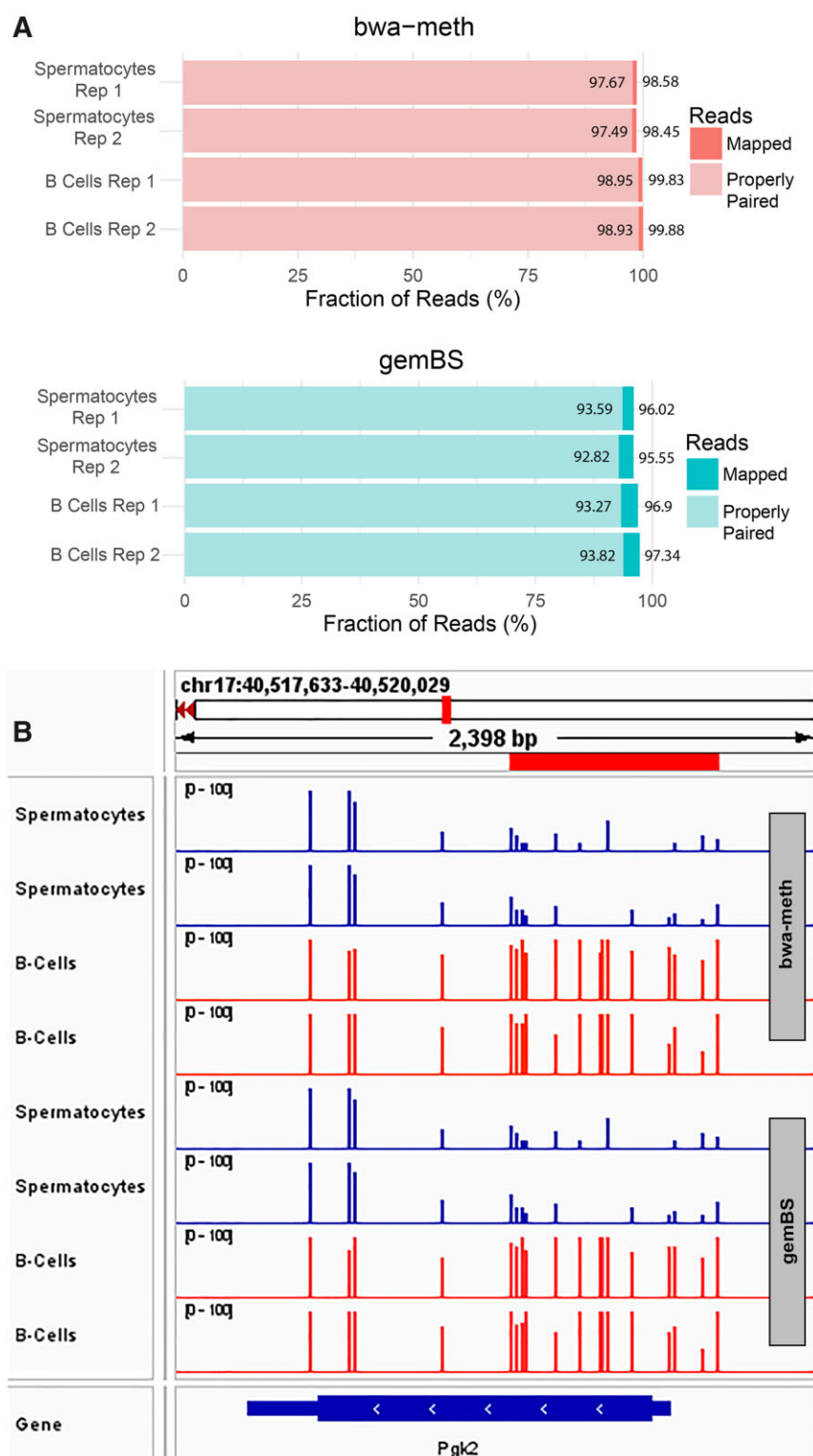


**Figure 2:** Comparison of alignment time in hours (hrs) when bwa-meth and gemBS aligners are used with WGBS FASTQ files ranging from  $8.5 \times 10^5$  to  $1.6 \times 10^8$  reads in size. As read counts increase there is a large difference in the time required for the bwa-meth and gemBS aligners to align FASTQ files. Comparison of the slopes of these rates indicates a time savings of  $0.71 \text{ h}/1 \times 10^7$  reads for when gemBS is used as the aligner instead of bwa-meth.

with gemBS requiring an average of only 1.43 h per file, whereas bwa-meth required an average of 11.36 h per file (Fig. 2). This indicates that gemBS increased alignment efficiency by  $0.71 \text{ h}/10^7$  reads, which translates to a more than  $\times 7$  improvement in the speed of sequence alignment for files containing 80–160 M reads when the gemBS aligner is used relative to that achieved by the bwa-meth aligner. As such, the gemBS aligner appears to provide greater utility for analysis of standard WGBS data.

When comparing the accuracy of the produced alignment files to determine if any increase in alignment speed afforded by the use of gemBS might be accompanied by reduced accuracy, we found the mapping accuracies among the three simulated read groups produced by the mason2 application were nearly identical. Specifically, we observed only minor differences in the number of reads that showed either failing mapping QC scores or paired reads mapping to different chromosomes, and in those contexts, gemBS displayed superior accuracy (Supplementary Section S4). To additionally test the alignment accuracy in a biological context, we compared the pipelines utilizing each aligner software for their relative ability to correctly identify a known DMR present in the *Pgk2* gene in mouse spermatocytes. We found that both bwa-meth and gemBS aligners were able to accurately map reads (Fig. 3A and Supplementary Section S2) and identify the DMR (Fig. 3B) in mouse spermatocytes when compared with mouse somatic CD19+ B cells, illustrating that replacement of the bwa-meth aligner with the gemBS aligner did not reduce the accuracy with which known DMRs previously identified by gene-specific analysis within a biological context are correctly revealed computationally by analysis of genome-wide WGBS data.

However, while these data suggest that the increased alignment speed afforded by gemBS is not associated with any reduction in the accuracy of identification of an expected DMR revealed by genome-wide WGBS analysis in mouse samples, we did observe small differences in the total number of other identified methylated regions including DMRs, UMRs/LMRs, and PMDs at the whole



**Figure 3:** Mapping accuracy and identification of a known DMR in spermatocytes compared with somatic B cells. **(A)** Comparison of the percentages of mapped and properly paired reads from spermatocyte and B-cell WGBS samples comparing the accuracy between bwa-meth and gemBS indicates that both aligners display a similar overall alignment accuracy. **(B)** Visualization of DNA methylation in the promoter region of the *Pfk2* gene in spermatocytes and B cells produced from either the bwa-meth or gemBS aligners shows that the accuracy of read mapping when either is used is sufficient to identify a DMR known to be present at the *Pfk2* promoter in spermatocytes when compared with somatic cell types.

epigenome level (Supplementary Section S5). For DMRs, we found that bwa-meth identified a total of 220 185 DMRs, while gemBS identified 212 045 DMRs with a significant portion ( $P$ -value =  $2.2e-16$ ) of each group containing DMRs that did not

overlap based on Pearson's Chi-squared test (10.76% and 7.34%, respectively). We found that just over half of these uniquely identified DMRs occurred in non-coding intergenic regions (58–60%). When comparing these unique non-overlapping DMRs identified by each

aligner, we found the median distance to the next nearest neighboring DMR produced by the alternative aligner was 2–3 kb. We also considered the possibility that the decrease in the overall number of DMRs identified by gemBS was due to the individual DMRs being longer. However, the median size of the bwa-meth and gemBS DMRs was similar (448 and 442 bp, respectively). The gemBS aligner identified a slightly larger number of predicted UMRs/LMRs, but slightly fewer predicted PMDs than the bwa-meth aligner in the data from nearly all mouse spermatocytes and CD19+ B cells tested. However, these differences were not found to be significant based on Kruskal–Wallis rank sum tests (UMRs/LMRs  $P$ -value = 0.39 and PMDs  $P$ -value = 0.77). Finally, the median of all segmented regions that were non-overlapping from each aligner was 9.73%, which was similar to the proportion found within DMRs, but differed in overall median distance of 23 kb to the next nearest regions identified by the opposing aligner. Thus, overall, there was strong agreement in the patterns of methylated regions identified from reads aligned by either aligner (~90%). However, some differences were noted in non-coding regions that could potentially be involved in transcriptional regulation.

## Discussion

The simplicity of the wg-blimp installation emanates from the use of Bioconda for package management during installation [52]. This allows wg-blimp to integrate several published software packages seamlessly into a single working environment, requiring only minimal technical expertise in software installation. This avoids the limited versatility associated with many pipelines that can restrict their integration into local computing systems. To optimally integrate gemBS into the improved pipeline, we updated the gemBS package on Bioconda to the most current version to overcome certain issues with its installation and use in a python environment that existed previously (Supplementary Section S6). This effort was important to maintain the overall simplicity of the wg-blimp installation process, as forcing users to compile gemBS manually could have negatively impacted the gain in overall pipeline speed we accomplished.

The increased speed of the alignment step afforded by replacing the bwa-meth aligner with the gemBS aligner represents a significant advance in the utility of the wg-blimp pipeline for analysis of WGBS or similar genome-wide DNA methylation data. When the bwa-meth aligner was used in conjunction with the wg-blimp pipeline, a full work week of computing time was normally required to complete an analysis of two sets of WGBS data each representing three replicates of samples sequenced to 80–160 M reads. However, replacement of the bwa-meth aligner with the gemBS aligner within the wg-blimp pipeline reduced the computing time required to accomplish the same procedure to a single day. In turn, this decrease in overall computing time strengthens the stability and utility of the pipeline by significantly reducing the potential for it to crash during long runs over multiple days, thus avoiding limitations imposed by computing networks that limit job times on nodes. An additional benefit of the gemBS aligner is that it automatically sorts the order of reads by chromosome in the output BAM files, whereas accomplishing this with the bwa-meth aligner requires an additional step in the wg-blimp pipeline. Finally, the gemBS aligner is better positioned to be adapted to rapid advancements in sequencing technologies, especially those applied to libraries with larger insert sizes and/or to higher read depths [53, 54].

Importantly, the increased analysis speed afforded by replacing the bwa-meth aligner with the gemBS aligner did not reduce

the overall analysis accuracy of the wg-blimp computational pipeline when tested on publicly available WGBS and EM-seq data. There were minor differences in the number of reads with low mapping QC scores, as well as in the number of paired reads mapping to different chromosomes in the BAM files, both of which are indications of reduced alignment accuracy. In both cases, the pipeline using the gemBS aligner actually performed better than that using the bwa-meth aligner. Thus, for these parameters, the gemBS-containing pipeline was more accurate than the bwa-meth-containing pipeline. Indeed, this exemplifies an additional limitation of the bwa-meth aligner when a seed has an exact match that occurs in multiple different chromosomes. To avoid more complex computational tasks to rule out all but one of the possible loci, the bwa-meth algorithm picks one of the chromosomes at random resulting in a higher rate of reads mapping to different chromosomes. Grouping of reads into different strata and completing searches through the reference genome by gemBS lower the overall number of reads where this occurs, giving the gemBS aligner another advantage over the bwa-meth aligner. Ultimately, both aligners are highly accurate, aligning nearly 100% of the reads supplied, as exemplified by the correct identification of the known DMR in the *Pgk2* gene promoter region when spermatocytes are compared with somatic cells. Thus, we conclude that there is no decline in alignment accuracy associated with the significant improvement in alignment speed afforded by use of gemBS aligner.

Despite finding differences in the numbers of all other methylated regions identified from reads aligned by either the bwa-meth or gemBS aligners, the close proximity of DMRs uniquely identified by the opposing aligner and the lack of statistical significance when comparing differences in the overall number of identified UMRs/LMRs and PMDs further indicate the two aligners yield very similar overall results, with the only potential differences occurring in non-coding intergenic regions which could be relevant to transcriptional regulatory sequences. To this end, our inclusion of the ability to toggle between use of either the original bwa-meth or newly added gemBS aligner in the wg-blimp pipeline is advantageous in that it affords the opportunity to focus on methylated regions that are conserved between both aligners to build conclusions based on identification of overlapping methylated regions. Ultimately, as noted above, the very significant increase in the computing speed afforded by inclusion of the gemBS aligner in the wg-blimp pipeline represents a substantial advance in the utility of this pipeline with only minor, if any, significant differences in the identification of DMRs within the epigenome.

## Conclusion

Replacement of the bwa-meth aligner with the gemBS aligner increased the overall speed of the alignment step in the wg-blimp pipeline by more than  $\times 7$  while maintaining high-level accuracy. This robustly increases the utility of this computational pipeline for analysis of WGBS or related genome-wide DNA methylation data. This modification removes one of, if not the only, source of concern about the previous version of the wg-blimp pipeline. With inclusion of the gemBS aligner, the wg-blimp pipeline represents a comprehensive, accurate, and rapid approach to the analysis of whole-genome DNA methylation data which can be utilized in a local core computing environment. In addition, this improvement positions the wg-blimp pipeline to be adaptable to future advancements in sequencing technologies or chemistries which will lead to libraries containing longer insert reads that

could be sequenced at much higher depths. As sequencing costs continue to decline and an increasing number of labs adopt whole-genome assessment of DNA methylation patterns as a commonly used epigenomic profiling assay, modifications of the sort reported here that significantly enhance the utility of specific analytical approaches will advance the field of epigenomic profiling. Thus, we believe that these changes to the wg-blimp pipeline will further ease the burden of processing epigenomic data associated with genome-wide DNA methylation patterns that accompany WGBS to help strengthen the field of epigenetic research.

## Declarations

### Data and code accessibility

These improvements have been merged with the wg-blimp source code as the newly updated v0.10.0 and are available at the following GitHub repository <https://github.com/MarWoe/wg-blimp> and can also be found on the wg-blimp Code Ocean compute capsule DOI 10.24433/CO.7135892.v1. The WGBS sperm and blood sample FASTQ datasets analyzed during the current study are available at <https://uni-muenster.sciebo.de/s/7vpqRSEATYcVlnP>. The mouse spermatocyte and CD19+ B-cell WGBS datasets and the human WGBS and EM-seq datasets created from DNA in blood samples collected from cancer patients analyzed during the current study are in the Gene Expression Omnibus, under accession numbers GSE161458, GSE49624, and GSE208549. The simulated read files generated by mason2 and analyzed during the current study can be generated by following the steps listed in [Supplementary Section S1](#), but are also available from the corresponding author on request. Additional information about the config file, commands, reference genome and annotation files, CpG island annotation file and repeat masker file used to run the wg-blimp pipeline during the current study can be found at <https://github.com/MarWoes/wg-blimp/issues/5>.

## Supplementary data

[Supplementary data](#) are available at *Biology Methods and Protocols* online.

## Acknowledgments

The authors would like to thank Conrad Weidenkeller for technical advice about environmental setup and troubleshooting while working in our HPC. They would like to thank Dr Marius Wöste for his advice while setting up the original wg-blimp pipeline and troubleshooting while adapting the pipeline to be run with mouse samples. In addition, they would like to thank Dr Wöste for merging our changes into the wg-blimp pipeline as v0.10.0. They would also like to thank Dr Simon Heath for trusting us with updating the gemBS software to v3.5.5\_IHEC on the Bioconda channel of the Anaconda package repository and Dr Devon Ryan for his initial help updating the gemBS Bioconda build from v3.2.0 to v3.5.0 as well as his tips and approval to pull requests of subsequent updates. Finally, they would like to thank Dr Yufeng Wang for reading the manuscript and providing useful suggestions. This work received computational support from UTSA's HPC cluster Arc, operated by Tech Solutions.

## Author contributions

Jake D. Lehle (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]) and John R. McCarrey (Formal analysis [supporting], Funding acquisition [lead], Project administration [lead], Resources [lead], Supervision [lead], Writing—review and editing [equal])

## Funding

This project was funded by the Robert J. and Helen C. Kleberg Foundation, the Nancy Hurd Smith Foundation and the following NIH grants to J.R.M.: NICHD P50 HD98593 and NIDA U01DA054179.

## Conflict of interest statement

None declared.

## References

- Moore LD, Le T, Fan G. DNA methylation and its basic function. *Neuropsychopharmacology* 2013;**38**:23–38.
- Gu H, Smith ZD, Bock C et al. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* 2011;**6**:468–81.
- Taiwo O, Wilson GA, Morris T et al. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* 2012;**7**:617–36.
- Fatemi M, Pao MM, Jeong S et al. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic Acids Res* 2005;**33**:e176.
- Beagan JA, Phillips-Cremens JE. On the existence and functionality of topologically associating domains. *Nat Genet* 2020;**52**:8–16.
- Zhou L, Ng HK, Drautz-Moses DI et al. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Sci Rep* 2019;**9**:1–16.
- Muir P, Li S, Lou S et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* 2016;**17**:1–9.
- Li M, Zou D, Li Z et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res* 2019;**47**:D983–8.
- Song Q, Decato B, Hong EE et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* 2013;**8**:e81148.
- Hackenberg M, Barturen G, Oliver JL. NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. *Nucleic Acids Res* 2011;**39**:D75–9.
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**:2520–2.
- Wöste M, Leitão E, Laurentino S et al. Wg-blimp: an end-to-end analysis pipeline for whole genome bisulfite sequencing data. *BMC Bioinformatics* 2020;**21**:8.
- Bhardwaj V, Heyne S, Sikora K et al. snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics* 2019;**35**:4757–9.
- Wurmus R, Uyar B, Osberg B et al. PiGx: reproducible genomics analysis pipelines with GNU Guix. *Gigascience* 2018;**7**:1–14.



15. Sun K, Li L, Ma L *et al.* Msuite: a High-performance and versatile DNA methylation data-analysis toolkit. *Patterns (New York, NY)* 2020;**1**:100127.
16. Kretzmer H, Otto C, Hoffmann S. BAT: bisulfite analysis toolkit. *F1000Research* 2017;**6**:1490.
17. Graña O, López-Fernández H, Fdez-Riverola F *et al.* Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. *Bioinformatics* 2018;**34**:1414–5.
18. Kim H, Sim M, Park N *et al.* msPIPE: a pipeline for the analysis and visualization of whole-genome bisulfite sequencing data. *BMC Bioinformatics* 2022;**23**:13.
19. Davis CA, Hitz BC, Sloan CA *et al.* The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2018;**46**:D794–801.
20. Ewels PA, Peltzer A, Fillinger S *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* 2020;**38**:276–8.
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
22. Merkel A, Fernández-Callejo M, Casals E *et al.* gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. *Bioinformatics* 2019;**35**:737–42.
23. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. <https://doi.org/10.48550/arXiv.1303.3997> (8 February 2023, date last accessed).
24. Marco-Sola S, Sammeth M, Guigó R *et al.* The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* 2012;**9**:1185–8.
25. Broad Institute. Picard Toolkit. <https://broadinstitute.github.io/picard/> (8 February 2023, date last accessed).
26. Andrews S, Krueger F, Segonds-Pichon A *et al.* Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (8 February 2023, date last accessed).
27. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 2016;**32**:292–4.
28. Ewels P, Magnusson M, Lundin S *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8.
29. Ryan D. dpryan79/MethylDackel: a (mostly) universal methylation extractor for BS-seq experiments. <https://github.com/dpryan79/MethylDackel> (8 February 2023, date last accessed).
30. Jühling F, Kretzmer H, Bernhart SH *et al.* metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res* 2016;**26**:256–62.
31. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;**13**:R83.
32. Schröder C. *Bioinformatics from Genetic Variants to Methylation*; 2018. <https://dx.doi.org/10.17877/DE290R-19925> (8 February 2023, date last accessed).
33. Burger L, Gaidatzis D, Schübeler D *et al.* Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res* 2013;**41**:e155.
34. Martin FJ, Amode MR, Aneja A *et al.* Ensembl 2023. *Nucleic Acids Res* 2023;**51**:D933–41.
35. Nassar LR, Barber GP, Benet-Pagès A *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* 2023;**51**:D1188–95.
36. Smit AFA, Hubley R, Green P. RepeatMasker Home Page. <https://www.repeatmasker.org/> (8 February 2023, date last accessed).
37. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* 2018;**34**:867–8.
38. Shiny CJ. <https://shiny.rstudio.com/> (8 February 2023, date last accessed).
39. Laurentino S, Cremers J-F, Horsthemke B *et al.* Healthy ageing men have normal reproductive function but display germline-specific molecular changes. *medRxiv* 2019;19006221.
40. Shukla V, Samaniego-Castruita D, Dong Z *et al.* TET deficiency perturbs mature B cell homeostasis and promotes oncogenesis associated with accumulation of G-quadruplex and R-loop structures. *Nat Immunol* 2022;**23**:99–108.
41. Hammoud SS, Low DHP, Yi C *et al.* Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell* 2014;**15**:239–53.
42. Holtgrewe M. Mason—a read simulator for second generation sequencing data. <https://publications.imp.fu-berlin.de/962/> (8 February 2023, date last accessed).
43. Li H, Handsaker B, Wysoker A *et al.*; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
44. McCarrey JR, Thomas K. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature* 1987;**326**:501–5.
45. Danshina PV, Geyer CB, Dai Q *et al.* Phosphoglycerate kinase 2 (PGK2) is essential for sperm function and male fertility in mice. *Biol Reprod* 2010;**82**:136–45.
46. Geyer CB, Kiefer CM, Yang TP *et al.* Ontogeny of a demethylation domain and its relationship to activation of tissue-specific transcription. *Biol Reprod* 2004;**71**:837–44.
47. McCarrey JR, Geyer CB, Yoshioka H. Epigenetic regulation of testis-specific gene expression. *Ann N Y Acad Sci* 2005;**1061**:226–42.
48. Füllgrabe J, Gosal WS, Creed P *et al.* Simultaneous sequencing of genetic and epigenetic bases in DNA. *Nat Biotechnol* 2023.
49. Kint S, De Spiegelaere W, De Kesel J *et al.* Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. *PLoS One* 2018;**13**:e0199091.
50. Schilbert HM, Rempel A, Pucker B. Comparison of read mapping and variant calling tools for the analysis of plant NGS data. *Plants* 2020;**9**:439.
51. King DJ, Freimanis G, Lasecka-Dykes L *et al.* A systematic evaluation of high-throughput sequencing approaches to identify low-frequency single nucleotide variants in viral populations. *Viruses* 2020;**12**:1187.
52. Dale R, Grüning B, Sjödin A *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;**15**:475–6.
53. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. *Front Genet* 2019;**10**:426.
54. Ou S, Liu J, Chougule KM *et al.* Effect of sequence depth and length in long-read assembly of the maize inbred NC358. *Nat Commun* 2020;**11**:1–10.