Accelerating the Alignment Processing Speed of the Comprehensive End-to-End Whole-Genome Bisulfite Sequencing Pipeline, wg-blimp

Jake D. Lehle (jake.lehle@utsa.edu)
The University of Texas at San Antonio
John R. McCarrey
The University of Texas at San Antonio

Research Article

Keywords: Whole-genome bisulfite sequencing, Analysis pipeline, Epigenetics, DNA methylation

Posted Date: June 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1666741/v1

Abstract

Background

Analyzing whole-genome bisulfite sequencing (WGBS) datasets is a time-intensive process due to the complexity and size of the input raw sequencing files and lengthy read alignment step during downstream data processing. This is particularly challenging with WGBS data because the conversion of all unmethylated Cs to Ts genome-wide renders read alignment a cumbersome computational process that can take up to a full work week of computing time. The objective of the study described here was to modify the read alignment algorithm associated with the wg-blimp pipeline to shorten the time required to complete this phase while retaining overall read alignment accuracy.

Results

Here we report improvements upon the recently published pipeline wg-blimp (whole-genome bisulfite sequencing methylation analysis pipeline) 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 > 7x acceleration in the processing speed of samples when scaled to larger publicly available FASTQ datasets containing 80–160 million (M) reads. Importantly, this acceleration was achieved while maintaining nearly identical accuracy of properly mapped reads when compared to data from the original pipeline.

Conclusion

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 pipeline that can produce high-quality data much more rapidly without compromising read accuracy.

Background:

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 limit 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), 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, 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 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 regions (DMRs), and methylation segmentation for profiling of DNA methylation states at regulatory elements [12]. 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 sequencing data, implementing the wg-blimp pipeline in its current form often requires extended computing time emanating from the conversion of unmethylated cytosines to uracils in the original DNA strand following bisulfite treatment. During PCR amplification these uracils are replaced with thymine, ultimately resulting in the conversion of C-G base pairs into T-A base pairs. Because most cytosines 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 [13] and gemBS [14], designed specifically for mapping bisulfite-converted reads, perform a two-stage mapping process. Cytosines on read 1 are fully converted to thymines while guanines on read 2 are fully converted to adenines. The reads are then aligned to either of two reference genomes where either all of the cytosines have been converted to thymines or all guanines have been converted to adenines. 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 conversion and alignment of all reads to multiple indexed genomes, followed by conversion back to the starting read sequence, the alignment step imposes a very time-consuming computational step.

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 [13, 15]. 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 filtering out, by default, notably shorter chains contained within longer chains (which are at least 50% and 38 bp shorter than the longer chain) [15]. 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 [15]. While these strategies have increased the potential size of the read that can be aligned using the BWA-MEM software up to 100 bp, the aligner that the gemBS software is built on, GEM3, allows for mapping lengths of up to 1 kb which can scale more quickly to large sequencing analyses while maintaining equal if not superior read mapping accuracy when compared to BWA-MEM [14]. This superiority largely comes from gemBS performing the conversion of read steps before and after mapping “on the fly” [14] 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 speed and accuracy over BWA-MEM and other heuristic mapping algorithms [16]. Searching through such a large index file does expose one limitation of gemBS, which is that it requires 48 GB of RAM compared to 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 [14]. 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 > 7x and open up the pipeline to the alignment of longer reads, all without sacrificing alignment accuracy.

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 [14, 17, 18]. 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). To accomplish this, we first determined the runtime required for alignment by each pipeline using existing small sample datasets provided to test the wg-blimp pipeline installation, which included isogenic human blood and sperm WGBS sequencing files (each generated from pools of DNA from six men) with nearly 1M reads each, all restricted to chr22 [12, 19]. We found the relative alignment speed was similar for both bwa-meth and gemBS with the average bwa-meth alignment time being only slightly shorter than the gemBS alignment time.

We next examined how these aligners performed when handling larger sequencing files. 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 with gemBS requiring an average of only 1.43 hours per file whereas bwa-meth required an average of 11.36 hours per file (Fig. 2). This indicates that gemBS increased alignment efficiency by 0.71 hrs/10^7 reads, which translates to a > 7x 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.

We then tested the extent to which this 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 [20] (Supplementary Sec. 1) with the number of reads in each group increasing by a log scale from 1-100 M reads. We defined accuracy as the percentages of both mapped and properly paired reads indicated in the BAM file output for each aligner using the flagstat function from SAMtools [21]. 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. The mapping accuracies among the three groups were nearly identical with 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 Sec. 2).

To further assess the accuracy of these aligners in a biological context we obtained publicly available mouse WGBS data from CD19 + B-cells [22] and spermatocytes [23] and used each aligner to analyze DNA methylation patterns in each dataset. We found both bwa-meth and gemBS aligners were able to accurately map reads (Fig. 3a) and identify a known DMR present in the *Pgk2* gene promoter region in spermatocytes (Fig. 3b). *Pgk2* is an intronless gene that arose via retrotransposition of the *Pgk1* gene and is required during normal spermatogenesis [24, 25]. Our lab has previously shown that in mice the upstream half of the *Pgk2* gene becomes demethylated in prospermatogonia and spermatogonia, prior to activation of transcription of the *Pgk2* gene in spermatocytes [26, 27]. Thus, we tested the accuracy with which pipelines utilizing each aligner software were able to correctly identify this known DMR. Both pipelines accurately identified this known DMR in spermatocytes when compared to 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. This data suggests that the increased alignment speed afforded by gemBS is not associated with any reduction in the accuracy of DMRs revealed by genome-wide WGBS analysis. All analyses were executed on a server equipped with one Intel Cascade Lake CPU with 80 physical cores and 160 hypertheads or virtual cores, 394 GB of memory, and a CentOS 7 operating system.

Discussion:

The simplicity of the wg-blimp installation comes from the use of Bioconda for package management during installation [28]. 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 Sec. 3]. 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 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 each. 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 when the bwa-meth aligner is used 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 longer insert sizes and/or to higher read depths [29, 30].

Importantly, the increased analysis 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 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 complete searches through the reference genome by gemBS, lowers 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.

Therefore, 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 no associated deleterious effects that we have detected.

Conclusion:

Replacement of the bwa-meth aligner with the gemBS aligner increased the overall speed of the alignment step in the wg-blimp pipeline while maintaining high-level accuracy. This robustly increases the utility of this computational pipeline for analysis of WGBS 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 WGBS 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 and chemistries which 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 WGBS as a commonly used epigenomic profiling assay, modifications of the sort reported here that significantly enhance the utility of specific methodologies will advance the field of epigenomic profiling. We believe these changes to the wg-blimp pipeline will further ease the burden of data processing that accompanies WGBS to help strengthen the field of epigenetic research.

Abbreviations

WGBS
whole-genome bisulfite sequencing
Wg-blimp
whole-genome bisulfite sequencing methylation analysis pipeline
DMR
differentially methylated region
SMEM
super-maximal exact match
QC
quality control
M
million

Declarations

Availability of data and materials

The improved wg-blimp source code is available at the following forked GitHub repository https://github.com/JakeLehle/wg-blimp. These changes have also been submitted as a pull request to

release a new version of wg-blimp v0.10.0.

The WGBS sperm and blood sample FASTQ datasets analyzed during the current study are available at https://uni-muenster.sciebo.de/s/7vpqRSEATYcvlnP. The spermatocyte and CD19+ B-cell WGBS datasets analyzed during the current study are in the Gene Expression Omnibus, under accession numbers GSE161458 and GSE49624.

The simulated read files generated by mason2 and analyzed during the current study are available from the corresponding author on reasonable request.

Additional information about the config file, commands, reference genome and annotation files, cgi 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.

Acknowledgments

This work received computational support from UTSA's HPC cluster Arc, operated by Tech Solutions. The authors would like to thank 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. The authors would also like to thank Heath Simon for trusting us with updating the gemBS software to v3.5.5_IHEC on the Bioconda channel of the Anaconda package repository, and Devan 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, the authors would like to thank Dr. Yufeng Wang for reading the manuscript and providing useful suggestions.

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 JRM: NICHD P50 HD98593, NIDA U01DA054179.

Author information

Affiliations

The University of Texas at San Antonio, 1 UTSA Circle, San Antonio, TX 78249

Jake D. Lehle and John R. McCarrey

Authors' Contributions

JDL updated the gemBS package on the Bioconda channel of the Anaconda package repository software, modified the wg-blimp software, and drafted the initial manuscript. JDL and JRM prepared the final manuscript. JRM provided funding for this project.

Corresponding author

Correspondence to Jake D. Lehle

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacol 2013 381. 2012;38:23–38.
2. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 2011 64. 2011;6:468–81.
3. Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, et al. Methylome analysis using MeDIP-seq with low DNA concentrations. Nat Protoc 2012 74. 2012;7:617–36.
4. Fatemi M, Pao MM, Jeong S, Gal-Yam EN, Egger G, Weisenberger DJ, 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.
5. Beagan JA, Phillips-Cremins JE. On the existence and functionality of topologically associating domains. Nat Genet 2020 521. 2020;52:8–16.
6. Zhou L, Ng HK, Drautz-Moses DI, Schuster SC, Beck S, Kim C, et al. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. Sci Reports 2019 91. 2019;9:1–16.
7. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biol. 2016;17:1–9.
8. Li M, Zou D, Li Z, Gao R, Sang J, Zhang Y, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. Nucleic Acids Res. 2019;47:D983–8.
9. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. PLoS One. 2013;8:81148.
10. Hackenberg M, Barturen G, Oliver JL. NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. Nucleic Acids Res. 2011;39 Database issue.

11. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.
12. Wöste M, Leitão E, Laurentino S, Horsthemke B, Rahmann S, Schröder C. Wg-blimp: An end-to-end analysis pipeline for whole genome bisulfite sequencing data. BMC Bioinformatics. 2020;21:1–8.
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754.
14. Merkel A, Fernández-Callejo M, Casals E, Marco-Sola S, Schuyler R, Gut IG, et al. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. Bioinformatics. 2019;35:737–42.
15. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.
16. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 2012 912. 2012;9:1185–8.
17. Schilbert HM, Rempel A, Pucker B. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants 2020, Vol 9, Page 439. 2020;9:439.
18. King DJ, Freimanis G, Lasecka-Dykes L, Asfor A, Ribeca P, Waters R, et al. A Systematic Evaluation of High-Throughput Sequencing Approaches to Identify Low-Frequency Single Nucleotide Variants in Viral Populations. Viruses. 2020;12.
19. Laurentino S, Cremers J-F, Horsthemke B, Tüttelmann F, Czeloth K, Zitzmann M, et al. Healthy ageing men have normal reproductive function but display germline-specific molecular changes. medRxiv. 2019;:19006221.
20. Holtgrewe M. Mason-A Read Simulator for Second Generation Sequencing Data FACHBEREICH MATHEMATIK UND INFORMATIK SERIE B • INFORMATIK. 2010.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
22. Shukla V, Samaniego-Castruita D, Dong Z, González-Avalos E, Yan Q, Sarma K, 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.
23. Hammoud SS, Low DHP, Yi C, Carrell DT, Guccione E, Cairns BR. Chromatin and Transcription Transitions of Mammalian Adult Germline Stem Cells and Spermatogenesis. Cell Stem Cell. 2014;15:239–53.
24. McCarrey JR, Thomas K. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature. 1987;326:501–5.
25. Danshina P V., Geyer CB, Dai Q, Goulding EH, Willis WD, Kitto GB, et al. Phosphoglycerate Kinase 2 (PGK2) Is Essential for Sperm Function and Male Fertility in Mice. Biol Reprod. 2010;82:136.
26. Geyer CB, Kiefer CM, Yang TP, McCarrey JR. Ontogeny of a Demethylation Domain and Its Relationship to Activation of Tissue-Specific Transcription1. Biol Reprod. 2004;71:837–44.

27. McCarrey JR, Geyer CB, Yoshioka H. Epigenetic Regulation of Testis-Specific Gene Expression. Ann N Y Acad Sci. 2005;1061:226–42.
28. Dale R, Grüning B, Sjödin A, Rowe J, Chapman BA, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 2018 157. 2018;15:475–6.
29. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. Front Genet. 2019;10 MAY:426.
30. Ou S, Liu J, Chougule KM, Fungtammasan A, Seetharam AS, Stein JC, et al. Effect of sequence depth and length in long-read assembly of the maize inbred NC358. Nat Commun 2020 111. 2020;11:1–10.

Figures

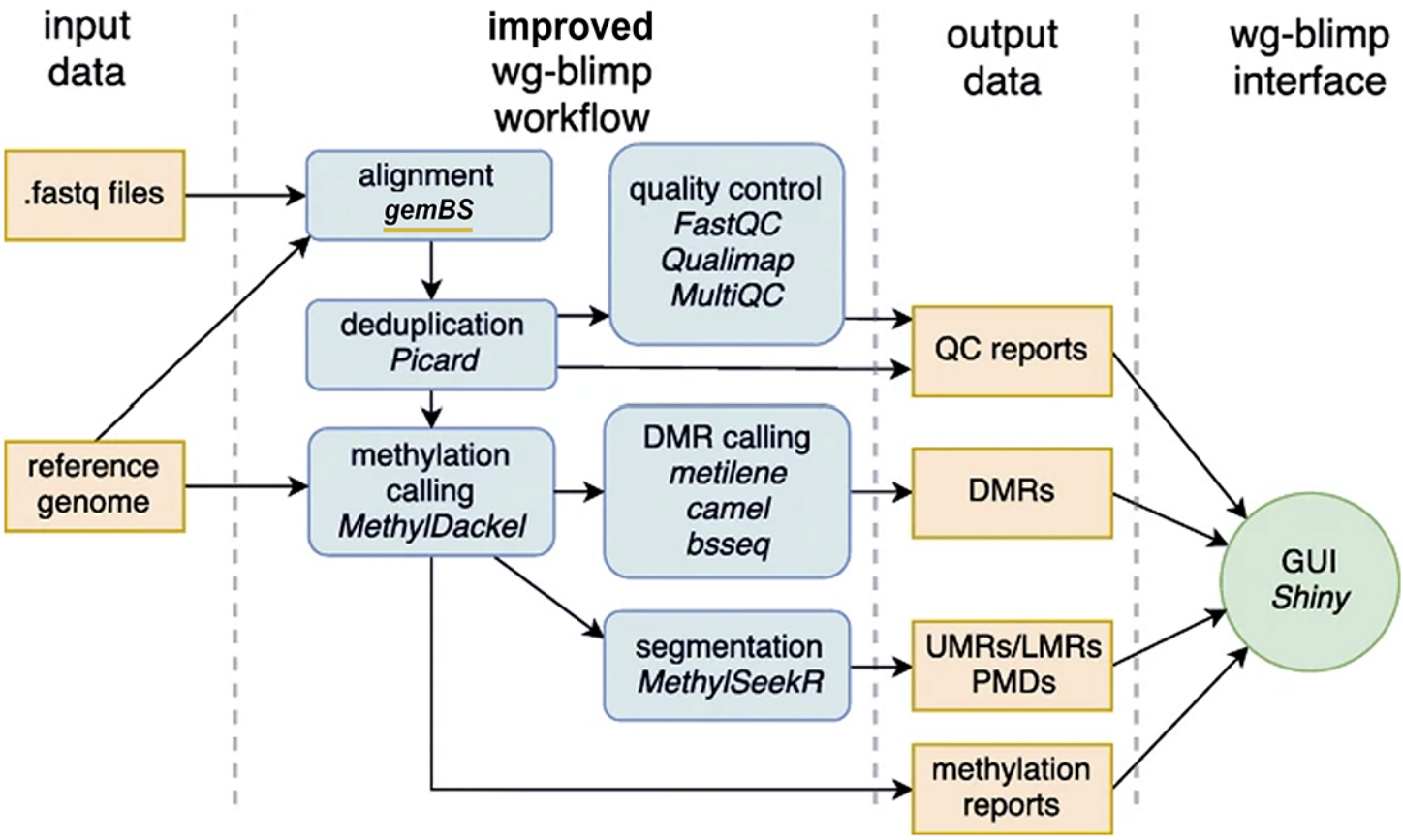

Figure 1

Improved wg-blimp workflow overview. FASTQ files and the reference genome file provided by the user are 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.

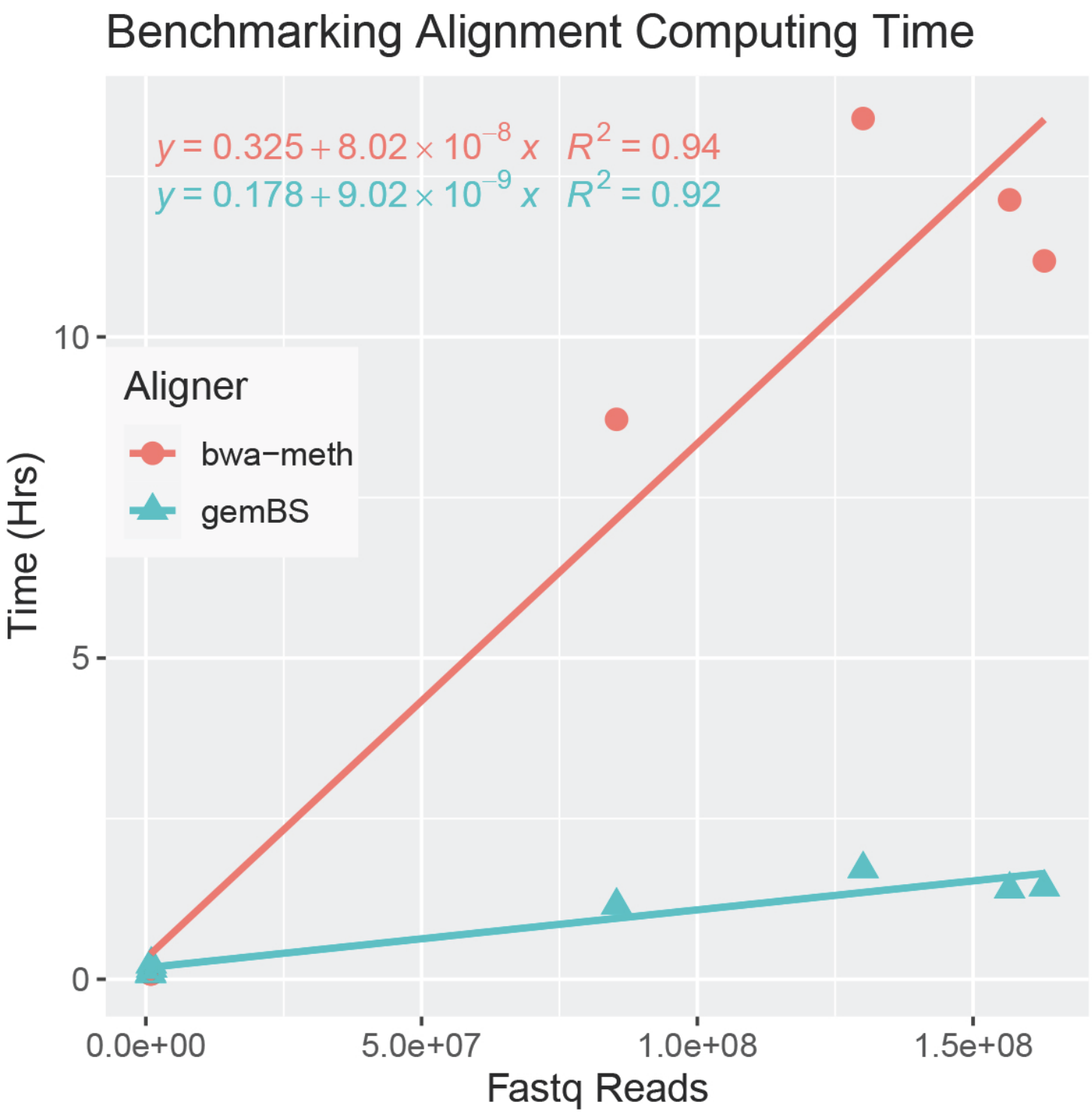

Figure 2

Comparison of alignment time in hours (hrs) when bwa-meth and gemBS aligners are used with WGBS FASTQ files ranging from 8.5x10^5 to 1.6x10^8 reads in size. As read counts increase there is a large difference in times 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 hrs/1x10^7 reads for gemBS over bwa-meth.

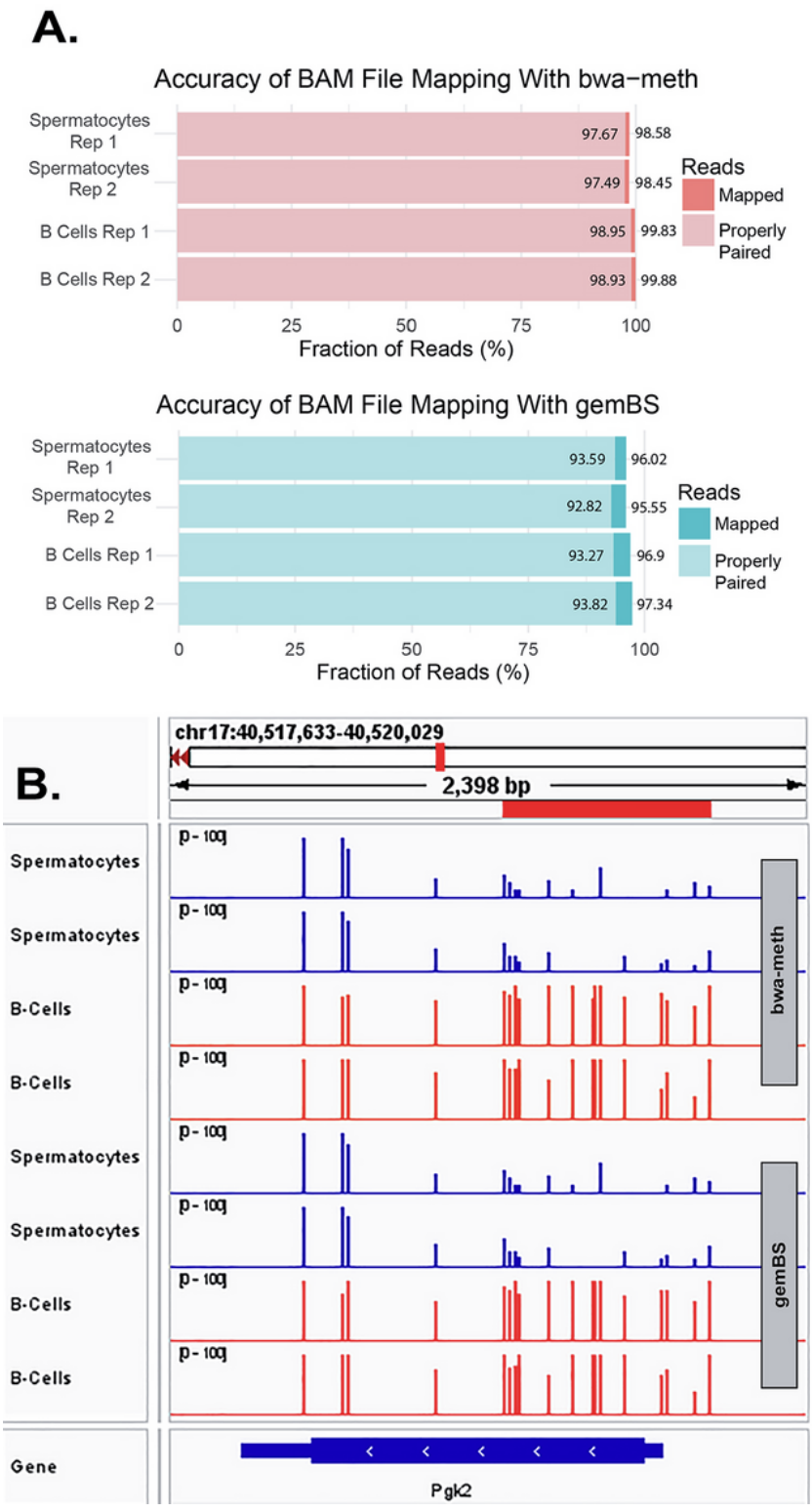

Figure 3

Mapping Accuracy and Identification of a Known DMR in Spermatocytes Compared to 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 aligners indicates that both aligners display a similar overall alignment accuracy. **B)** Visualization of DNA methylation in the promoter region of the *Pgk2* gene in spermatocytes and B-cells produced from either bwa-meth or gemBS

shows that the accuracy of read mapping when either aligner is used is sufficient to identify a DMR known to be present at the *Pgk2* promoter in spermatocytes when compared to somatic cell types.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Acceleratingthealignmentofwgblimpsupplementaryfinal.docx