GigaScience

# BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | GIGA-D-20-00299R1 |
| **Full Title:** | BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform |
| **Article Type:** | Technical Note |
| **Funding Information:** | National Institutes of Health (T32CA201160) — Mr. Colin Farrell |
| **Abstract:** | Abstract<br>Background: Bisulfite sequencing is commonly employed to measure DNA methylation. Processing bisulfite sequencing data is often challenging due to the computational demands of mapping a low complexity, asymmetrical library and the lack of a unified processing toolset to produce an analysis ready methylation matrix from read alignments. To address these shortcomings, we have developed BiSulfite Bolt (BSBolt); a fast and scalable bisulfite sequencing analysis platform. BSBolt performs a pre-alignment sequencing read assessment step to improve efficiency when handling asymmetrical bisulfite sequencing libraries.<br>Findings: We evaluated BSBolt against simulated and real bisulfite sequencing libraries. We found that BSBolt provides accurate and fast bisulfite sequencing alignments and methylation calls. We also compared BSBolt to several existing bisulfite alignment tools and found BSBolt outperforms Bismark, BSSeeker2, BISCUIT, and BWA-Meth based on alignment accuracy and methylation calling accuracy.<br>Conclusion: BSBolt offers streamlined processing of bisulfite sequencing data through an integrated toolset that offers support for simulation, alignment, methylation calling, and data aggregation. BSBolt is implemented as a python package and command line utility for flexibility when building informatics pipelines. BSBolt is available at https://github.com/NuttyLogic/BSBolt under an MIT license. |
| **Corresponding Author:** | Matteo Pellegrini<br>University of California Los Angeles<br>Los Angeles, CA UNITED STATES |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author's Institution:** | University of California Los Angeles |
| **Corresponding Author's Secondary Institution:** | |
| **First Author:** | Colin Farrell |
| **First Author Secondary Information:** | |
| **Order of Authors:** | Colin Farrell<br>Michael Thompson<br>Anela Tosevska<br>Adewale Oyetunde<br>Matteo Pellegrini |
| **Order of Authors Secondary Information:** | |
| **Response to Reviewers:** | Dear Editor Zhou,<br><br>We wish to thank the reviewers for their useful comments. We have revised the manuscript to address the comments. Please find our point-by-point responses below, marked in blue.<br><br>Reviewer #1: |

| | |
|---|---|
| | The authors introduce BSBolt as a complete pipeline for processing bisulfite sequence data.The main stated contribution of this tool over the authors' previous work of BSSeeker and BSSeeker2 is the direct modification of BWA-mem to align BS-Seq reads directly which results in increased accuracy. This seems like a nice improvement over existing methods. Several clarifications/changes in the paper would be helpful.<br><br>1. Given that the authors have previously written BSSeeker and recently written BSSeeker2,<br>more concise motivation of what short-coming BSBolt addresses in specifically those two<br>tools would be helpful.<br><br>Response:<br>We expanded on the motivation to develop BSBolt in the manuscript text below. In summary, we had several goals with BSBolt. We sought to improve performance, eliminate the alignment constraints enforced by BSSeeker / BSSeeker2, preserve the read level methylation calling and read bisulfite conversion checks, and finally we wanted to take a minimal performance hit for alignment of undirectional libraries. The three base alignment strategy as implemented by BSSeeker2 and Bismark has several limitations. Both tools carry out multiple intermediary alignments to separate alignment indices representing different reference conversion patterns and then integrate intermediate alignments together into a consensus alignment file. Reads with multiple alignments within an intermediate alignment file or across multiple intermediate alignment files are discarded; only reads that align uniquely within a single intermediate alignment are reported. In an effort to reduce the number of reads that align across alignment indices both BSSeeker2 and Bismark have strict default alignment parameters. In addition to being computationally demanding, this implementation can also reduce the number of valid alignments reported, as only the highest quality, unique alignments are output. BWA-Meth resolves this issue by performing alignment to a single bisulfite converted alignment index and processing reads on the fly; but, does not return the read level methylation calls or bisulfite conversion assessment provided by Bismark and BSSeeker2. Additionally, when performing bisulfite sequencing alignment the read conversion pattern is dependent on whether the sequenced DNA fragment is representative of the original DNA sequence or its PCR product. In a directional bisulfite sequencing library only DNA representative of the original DNA fragment is sequenced so the bisulfite conversion pattern is known. In an undirectional library, DNA representative of the original DNA fragment and its PCR product is sequenced so a cytosine to thymine or a guanine to adenine conversion is possible. BS-Seeker2 and Bismark handle undirecitonal libraries by converting input reads using both conversion patterns. This approach doubles the number of reads that must be aligned and generates input reads that will not be represented in the alignment index. BWA-Meth does not support alignment of undirectional libraries.<br><br>2. In my experience, installing pysam is more difficult than installing bwa-mem. So the statement "wrapping external read alignment tools introduces added complexity" is incorrect for example as it relates to bwa-meth. I expect the same is true for bismarck/bowtie(2). In addition to pysam, this tool seems to rely on samtools for methylation calling. That said, I was able to easily install this tool with pip.<br><br>Response:<br>We removed the statement about the added complexity of wrapping an alignment tool in the manuscript. Additionally, we have implemented Anaconda build recipes and added Anaconda installation instructions to the BSBolt documentation to provide a managed installation option.<br><br>3. This note:<br>"A read, or read pair, with a low proportion observed cytosines compared to guanine will be preferentially aligned<br>with a cytosine to thymine conversion pattern and vice versa. If it is unclear what conversion pattern<br>should be used, both conversion patterns are aligned and the conversion pattern with the highest total alignment score is output." indicates the most important algorithmic |

improvement in BSBolt. A sentence indicating this strategy in the abstract would motivate the tool early on. Also please include additional detail on the exact value f "low proportion"

Response:
We added text to the manuscript to explicitly state what proportion level is set to by default and what was used to the comparison. Additionally, we highlighted the use of the read assessment in the abstract and several additional places throughout the paper.
A read, or read pair, with a low proportion of observed cytosines compared to guanine (0.1 by default) will be preferentially aligned with a cytosine to thymine conversion pattern and vice versa.

4. What is the motivation for this: " Each alignment and methylation
calling workflow was given a maximum runtime of 24 hours. If an alignment was incomplete at the end of 24 hours, duplicate read marking and methylation calling was performed on the reads aligned during the 24 hour limit. " ?
It would be clearer to let each tool complete in whatever time it takes and then report the time along with the full results.

Response:
We provided each tool with a maximum run time of 288 hours and updated the text accordingly (below). All incomplete Bismark alignments completed in this time, but several BSSeeker2 alignments were unfinished after the 288 hour limit. We acknowledge the unfinished alignments are a limitation of the manuscript, but it wasn't feasible for us to extend the time limit beyond 288 hours.

5. Please add Table Legends
Response:

We added a table legend to table 1. Table 2 is represented as a figure in the revision and has captioned accordingly.

6. "The first 10kb of chr1 was duplicated and added as an additional contig."
This is all 'N' bases. What's the purpose of this?
Response:

The phrasing of the original sentence was incorrect. We duplicated the first 10kb of the simulated chr1. We clarified the text to highlight this.
The first 10kb of the simulated chr1 was duplicated and added as an additional contig.

7. nuttylogic.github.com/BSBoltManuscript is not available so I am not able to see the code to reproduce this analysis.
Likewise: https://bsbolt.readthedocs.io/ does not load (this might be an ephemeral clouflare issue).
I think this is the code used:
https://github.com/NuttyLogic/BSBoltManuscript/blob/master/AlignCompWGBS.py
In which case, if sam->bam conversion is used, it would be more fair to allow samtools view to use --threads if that is a bottleneck.

Response:

We evaluated if the samtools sam to bam conversion was a bottleneck for BISCUIT and BWA-METH. We also included BSBolt in the comparison as the conversion pipeline is built with htslib. There was performance gain when additional conversion threads were added. The addition of threads past 2 threads resulted in minimal performance gain so we set the conversion threads at 2. We also exposed the number of conversion threads as an option in BSBolt to prevent any bottlenecks for users. We now state this in the manuscript text and added a supplemental figure / text on this (text below).

Samtools and BSBolt were provided with two compression threads to minimize any alignment bottlenecks (supplemental figure 1).

8. In Table 2, please indicate that bwa-meth does not support undirectional and therefore the tool is not being used as intended.

Response:

We switched Table 2 with a figure in the revision and noted the BWA-Meth results accordingly.

Signed,
Brent Pedersen

Reviewer #2:
The authors present BSBolt, an analysis platform for processing bisulfite sequencing data. BSBolt introduces a new alignment file structure that allows rapid methylation calling. Benchmarking was performed using already existing tools, such as Bismark, BSSeeker2, BWA-Meth and BISCUIT. The BSBolt offers a very nice performance both in speed and accuracy.

Generally, the paper is well written, the results are clearly communicated. The BSBolt software is available with detailed documentation and a relatively easy installation. I needed to separately run 'make' for softwares in the External folder, maybe it worth mentioning in the documentation.
I have a few questions and suggestions:

Response:

We updated the github readme and documentation with more detailed installation instructions.

1. In the simulation, why did the authors use 0.05 as a mutation rate? If I interpret it correctly it is quite high, much higher than the general mutation rate for human. It might affect the performance of some tools, such as Bismark.

Response:

The mutation rate was set to 0.005 for the simulations, lower than 0.05, but it is certainly high compared to the expected human mutation rate and we expect around 0.5 genetic variants per 100bp. Directional reads simulated with a mutation rate of 0.005 and sequencing error rate of 0.005 were aligned accurately by both Bismark and BSSeeker2 (> 99.9%) with high mappability of 94.4% and 98.3% on average respectively (Table S1). Both Bismark and BSSeeker2 performed well at this baseline mutation rate. When the simulated error rates were increased to 0.02 BSSeeker2 and Bismark exhibited low mappability (Figure 2B) but the returned alignments were accurate (Table S1).

2. I was quite surprised by the low performance of Bismark. According to our experience, although slow and resource intensive, Bismark is quite accurate. In the simulation experiment the high mutation rate might explain this low performance, but it is the same with real data. Using similar computational setting, I don't recall Bismark taking us this long even with a somewhat bigger dataset. Did the authors check if the settings are adequate? The memory need increases quickly with the number of cores, can it be that it is limited by the amount of available memory? Using less cores might improve it. Are the accuracy results similar to those in the original publication about the dataset that was published (DOI:10.21203/rs.3.rs-33940/v1)? They also used Bismark there and compared it to Illumina array.

Response:
Bismark exhibited better performance when aligning the real data compared with the simulated read data. With the simulated directional, 100bp, paired-end reads Bismark performed the alignment in approximately 38 minutes (~35,500 reads / minute) compared with 3.239 minutes for BSBolt (~38X slower, table S2). With the real data

Bismark aligned the libraries in 71.1 hours on average (~29,000 read pairs / minute) compared with BSBolt in 3.61 hours on average (~20X slower, table S3). In terms of accuracy, Bismarck, BSBolt, BISCUIT, and BWA-Meth all exhibited accuracy in line with previously reported results and manuscript text was updated to reflect this (text below). Additionally, outside of BSSeeker2, all alignment tools showed low MAE between the sequencing methylation values and the array methylation values (Figure 3C).
The called methylation values were highly correlated with the sites called on the EPIC array across all alignment tools (Pearson's r=.92-98, supplemental table 2), as previously reported (Shu et al., 2020)

3. It would be interesting to see how BSBolt scales. What are the memory needs with 12 cores? Does it scale linearly? How fast it can be in a HPC environment with much more resources? It would be interesting to see a table or figure about it.

Response:

We added supplemental Figure 2 to show run time and memory consumption based on the number of alignment threads for single / paired end and directional / undirectional libraries 150bp libraries. Memory consumption increases linearly with the number of alignment threads. Run time decreases with added alignment threads, but the absolute run time is minimally changed by more threads after 8.

4. Despite it clearly shows good results, I think a more detailed rationale behind BSBolt would be nice, since BISCUIT offers very similar functionality with a slighter worse performance.

Response:

We expanded on the motivation to develop BSBolt in the manuscript text below. In summary, we had several goals with BSBolt. We sought to improve performance, eliminate the alignment constraints enforced by BSSeeker / BSSeeker2, preserve the read level methylation calling and read bisulfite conversion checks, and finally we wanted to take a minimal performance hit for alignment of undirectional libraries.
The three base alignment strategy as implemented by BSSeeker2 and Bismark has several limitations. Both tools carry out multiple intermediary alignments to separate alignment indices representing different reference conversion patterns and then integrate intermediate alignments together into a consensus alignment file. Reads with multiple alignments within an intermediate alignment file or across multiple intermediate alignment files are discarded; only reads that align uniquely within a single intermediate alignment are reported. In an effort to reduce the number of reads that align across alignment indices both BSSeeker2 and Bismark have strict default alignment parameters. In addition to being computationally demanding, this implementation can also reduce the number of valid alignments reported, as only the highest quality, unique alignments are output. BWA-Meth resolves this issue by performing alignment to a single bisulfite converted alignment index and processing reads on the fly; but, does not return the read level methylation calls or bisulfite conversion assessment provided by Bismark and BSSeeker2. Additionally, when performing bisulfite sequencing alignment the read conversion pattern is dependent on whether the sequenced DNA fragment is representative of the original DNA sequence or its PCR product. In a directional bisulfite sequencing library only DNA representative of the original DNA fragment is sequenced so the bisulfite conversion pattern is known. In an undirectional library, DNA representative of the original DNA fragment and its PCR product is sequenced so a cytosine to thymine or a guanine to adenine conversion is possible. BS-Seeker2 and Bismark handle undirecitonal libraries by converting input reads using both conversion patterns. This approach doubles the number of reads that must be aligned and generates input reads that will not be represented in the alignment index. BWA-Meth does not support alignment of undirectional libraries.

I also have some minor comments/recommendations:
I think table 2 would look better in a series of small figures, it would be quicker to go through the results.In the supplementary table 1, the "Aligned reads/min" should be "Million aligned reads/min".

| | |
|---|---|
| | Although python installation is easy, maybe it would worth making it available in conda or as a docker container for smoother integration in different environments.<br><br>Response:<br><br>We removed Table 2 and added a figure summarizing the results in its place. The labels for supplementary table 1 have been fixed. Additionally, we added conda build recipes for macOS and linux 64 to ease installation. |
| **Additional Information:** | |
| **Question** | **Response** |
| Are you submitting this manuscript to a special series or article collection? | No |
| **Experimental design and statistics**<br><br>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.<br><br>Have you included all the information requested in your manuscript? | Yes |
| **Resources**<br><br>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.<br><br>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist? | Yes |
| **Availability of data and materials**<br><br>All datasets and code on which the conclusions of the paper rely must be either included in your submission or | Yes |

| | |
|---|---|
| deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.<br><br>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist? | |

# BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform

Colin Farrell[1], Michael Thompson[2], Anela Tosevska[2], Adewale Oyetunde[2], Matteo Pellegrini[2,*]

1 Department of Human Genetics, University of California, Los Angeles, CA, USA;
2 Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA, USA;

***Corresponding Author**
Matteo Pellegrini
matteop@mcdb.ucla.edu

ORCID IDs:

Colin Farrell: 0000-0002-3138-6108
Michael Thompson: 0000-0002-2914-1080
Anela Tosevska: 0000-0002-0892-7068
Matteo Pellegrini:0000-0001-9355-9564

**Abstract**

**Background:** Bisulfite sequencing is commonly employed to measure DNA methylation. Processing bisulfite sequencing data is often challenging due to the computational demands of mapping a low complexity, asymmetrical library and the lack of a unified processing toolset to produce an analysis ready methylation matrix from read alignments. To address these shortcomings, we have developed BiSulfite Bolt (BSBolt); a fast and scalable bisulfite sequencing analysis platform. BSBolt performs a pre-alignment sequencing read assessment step to improve efficiency when handling asymmetrical bisulfite sequencing libraries.

**Findings:** We evaluated BSBolt against simulated and real bisulfite sequencing libraries. We found that BSBolt provides accurate and fast bisulfite sequencing alignments and methylation calls. We also compared BSBolt to several existing bisulfite alignment tools and found BSBolt outperforms Bismark, BSSeeker2, BISCUIT, and BWA-Meth based on alignment accuracy and methylation calling accuracy.

**Conclusion:** BSBolt offers streamlined processing of bisulfite sequencing data through an integrated toolset that offers support for simulation, alignment, methylation calling, and data aggregation. BSBolt is implemented as a python package and command line utility for flexibility when building informatics pipelines. BSBolt is available at https://github.com/NuttyLogic/BSBolt under an MIT license.

## Findings

## Background

DNA methylation, the epigenetic modification of cytosine by the addition of a methyl group to the fifth carbon of the cyclic backbone, is a widely studied epigenetic mark associated with gene regulation[1,2] and numerous biological processes [3–5]. High throughput sequencing combined with bisulfite conversion is a broadly used method for profiling DNA methylation genome wide[6][7]. Treatment of DNA with sodium bisulfite results in unmethylated cytosines being deaminated to uracil, and converted to thymine through PCR amplification, while methylated cytosine, guanine, thymine, and adenine remain unchanged [8]. The methylation status of an individual site or region can be assessed by looking at the number bisulfite converted bases relative to the total number of observed bases. Amongst eukaryotic organisms the majority of genomic cytosines are unmethylated [8–10]. As a consequence, bisulfite sequencing reads originating from the same location but opposite strands are generally no longer complementary. Additionally, when the PCR product of the original bisulfite converted sequence is considered, sequencing reads can be aligned in different orientations within the same strand. Given the asymmetrical nature of bisulfite sequencing libraries and the large number of potential mismatches between the read sequence and the reference the use of a traditional alignment tool would produce low quality alignments.

Bisulfite sequencing alignment tools Bismark[11], BS-Seeker2[11,12], and BWA-Meth[13] successfully adopted a three-base alignment strategy wrapped around established read aligners such as

Bowtie2[14,15] and BWA-MEM[14], to accurately align bisulfite sequencing reads. In this strategy, an alignment index or multiple alignment indices are generated against each bisulfite converted reference strand. Relative to the reference, the bisulfite sense strand is the reference with all cytosines converted to thymine and the antisense strand is the reference sequence with all guanines converted to adenine. Before alignment, input reads are *in silico* bisulfite converted so any methylated or incompletely converted bases are converted to remove mismatches relative to the bisulfite reference. Reads are then aligned using the wrapped read alignment tool and the output alignments are integrated together with the original read sequence to form a consensus alignment file. During the generation of a consensus alignment file BS-Seeker2 and Bismark call contextual methylation, where CG methylation is reported distinctly from CH (H=A,C,T) methylation, for every aligned base within an alignment. The regional methylation information provided within alignment calls can provide important context about the epigenetic organization of a genome and the reorganization that occurs in response to disease [16–18]. Methylation calls from aligned reads can also be leveraged to assess the bisulfite conversion status of a read. A high proportion of observed methylated CH sites relative to the total number of observed CH indicates a read that was incompletely bisulfite converted as the majority of CH sites are expected to be unmethylated.

The three base alignment strategy as implemented by BSSeeker2 and Bismark has several limitations. Both tools carry out multiple intermediary alignments to separate alignment indices representing different reference conversion patterns and then integrate intermediate alignments together into a consensus alignment file. Reads with multiple alignments within an intermediate alignment file or across multiple intermediate alignment files are discarded; only reads that align uniquely within a single intermediate alignment are reported. In an effort to reduce the number of reads that align across alignment indices both BSSeeker2 and Bismark have strict default alignment parameters. In addition to being computationally demanding, this implementation can also reduce the number of valid alignments reported, as only the highest quality, unique alignments are output. BWA-Meth resolves this issue by performing alignment to a single bisulfite converted alignment index and processing reads on the fly; but, does not return the read level methylation calls or bisulfite conversion assessment provided by Bismark and BSSeeker2. Additionally, when performing bisulfite sequencing alignment the read conversion pattern is dependent on whether the sequenced DNA fragment is representative of the original DNA sequence or its PCR product. In a directional bisulfite sequencing library only DNA representative of the original DNA fragment is sequenced so the bisulfite conversion pattern is known. In an undirectional library, DNA representative of the original DNA fragment and its PCR product is sequenced so a cytosine to thymine or a guanine to adenine conversion is possible. BS-Seeker2 and Bismark handle undirecitonal libraries by converting input reads using both conversion patterns. This approach doubles the number of reads that must be aligned and generates input reads that will not be represented in the alignment index. BWA-Meth does not support alignment of undirectional libraries.

Here we present BiSulfiteBolt (BSBolt), a bisulfite sequencing platform designed to be fast and scalable while also providing the same read-level methylation calls and quality metrics of BS-Seeker2 and Bismark to preserve compatibility with existing analysis tools. BSBolt alignment is built on a forked version of BWA-MEM[14,19] and HTSLIB[19] with bisulfite specific sequencing logic integrated directly into the alignment process. BSBolt incorporates a pre-alignment read assessment step to assess the correct conversion pattern when aligning undirectional libraries. This eliminates the needs to perform multiple alignments for the same read, improving performance. Additionally, as the output alignment structure is slightly different between each bisulfite alignment wrapper, each tool implements its own methylation calling utility and output format. BSBolt includes a rapid and multi-threaded methylation caller, that outputs methylation calls in CGmap or bedGraph format implemented by BSSeeker2 and Bismark respectively. We show that BSBolt alignments and methylation calling is considerably faster and more accurate than these other bisulfite sequencing alignment wrappers. Additionally, we compare BSBolt to another high performance bisulfite sequencing platform BISCUIT[20]**.** BISCUIT also incorporates bisulfite specific alignment logic directly into the alignment process, but doesn't support read level methylation calling or bisulfite conversion assessment during alignment. Despite this, we show that BSBolt offers comparable, or faster, performance. Additionally, to facilitate end to end processing of bisulfite sequencing data BSBolt includes utilities for read simulation utility and aggregation of methylation call files into a consensus matrix.

**Methods**:

**BSBolt Workflow**

BSBolt Alignment

BSBolt incorporates bisulfite alignment logic directly within a forked version of BWA-MEM. BSBolt is designed around a single Burrows-Wheeler Transform (BWT) FM-index constructed from both bisulfite converted reference strands. BSBolt utilizes a three base alignment strategy where input reads sequences are fully *in silico* converted before alignment. In this case of undirectional libraries, where a cytosine to thymine or guanine to adenine conversion if possible, BSBolt first analyzes the read base composition. A read, or read pair, with a low proportion of observed cytosines compared to guanine (0.1 by default) will be preferentially aligned with a cytosine to thymine conversion pattern and vice versa. If it is unclear what conversion pattern should be used, both conversion patterns are aligned and the conversion pattern with the highest total alignment score is output. The converted read sequence is aligned using BWA-MEM to the bisulfite FM-index. The resulting alignments are then modified so reads mapping to the sense reference strand are reported as sense reads and the anti-sense reference reported as antisense reads regardless of mapping orientation. The mapping quality of an alignment is assessed by mapping uniqueness using standard BWA-MEM scoring criteria. Additionally, an alignment with alternative alignments on a different bisulfite reference strand is further penalized for being bisulfite

ambiguous. Read variation and methylation calls are then made for alignments meeting scoring thresholds using the original read sequence and an unconverted reference sequence. If a difference between the alignment and reference is explainable by bisulfite conversion a methylation call is made for the aligned base; otherwise, reference variation is reported. When calling methylation values, the context of the methylatable base is considered by capturing the local reference context (ie CG or CH). The methylation calls are output as a Sequence Alignment/Map (SAM) flag mirroring the BWA-MEM MD flag. Typically, the majority of CH sites are unmethylated so the expectation is that the majority of CH sites within a read, or read pair, are bisulfite converted. After calling read level methylation this information is leveraged to assess the bisulfite conversion status of the read across all aligned bases within the read, or read pair. The conversion status of the read is conveyed as a SAM flag in the output alignment. Output alignments are then compressed and written to a bam file natively.

BSBolt Methylation Calling

BSBolt includes an optimized methylation calling utility that takes advantage of the BSBolt alignment file structure to rapidly call site methylation. The calling procedure proceeds as follows. A read pileup is created using SAMtools (SAMTOOLS, RRID:SCR_002105) [21], and initialized using pysam[22], for each reference contig with aligned reads. Methylation calls are made for all methylatable bases, or only CG sites, using all reads that pass user specified quality metrics. Methylation values for reference guanine nucleotides are made for reads aligned to the antisense strand and calls for reference cytosine nucleotides are made for reads aligned to the sense strand. This call strategy decreases methylation calling time, as information about the origin strand can be quickly interpreted. Methylation calls are then output in the CGmap file format implemented by BSSeeker2. To aggregate several call files together into a consensus matrix BSBolt includes a rapid and efficient matrix aggregation utility. Bisulfite sequencing techniques often capture methylation sites unevenly, so making a combined matrix of all sites observed across every call file can be inefficient and produce large sparse matrices. BSBolt utilizes an iterative matrix assembly method where individual CGmap files are iterated through to count how often individual sites appear at or above a user specified coverage threshold. If a site is observed in a set proportion of the CGmap files the site is included in the consensus matrix. This process is parallelizable across several threads for efficiency. BSBolt supports output of matrices containing methylation values and counts of methylated and total bases at each site.

BSBolt Simulation

BSBolt Simulate utilizes a modified version of WGSIM[23] wrapped with python to simulate bisulfite converted reads with site specific methylation information incorporated across reads. Given a reference sequence global methylation values are set by randomly selecting a methylation value for all methylatable bases depending on context (CG or CH) or by passing a methylation profile in the form of a CGmap file. Reads are then simulated by randomly selecting a genomic position within a reference

sequence, sampling the reference sequence at set read length, and insert size for paired end reads, then incorporating sequencing error and genetic variation. The origin strand, and conversion pattern if simulating undirectional reads, is then randomly selected. At every methylatable base within a read the methylation status of the base is set by the probability of observing a methylated base given the reference methylation value. The mapping location, methylation status, and origin bisulfite strand are attached as a fastq comment and output along with the bisulfite converted read sequence and base call qualities. The number of methylated and unmethylated bases covering each methylation site are output as a serialized python object at the end of the simulation.

**Tool Comparisons**

BSBolt (v1.4.4), BISCUIT (v0.3.16.20200420), BSSeeker2 (v2.1.8), BWA-Meth (v0.2.2), and Bismark (v0.22.3) were used for comparisons with both real and simulated bisulfite sequencing data. All comparisons were performed on a compute node with XEON X5650 six core (twelve thread) processor (48GB ram) running centos (v6.10). Each tool was provided with 12 compute threads if supported. Default alignment parameters were used unless library specific alignment options were necessary to support the simulated library type. Uncompressed alignment outputs were compressed using SAMtools (v1.9) before being written to disk [24]. SAMtools and BSBolt were provided with two compression threads to minimize any alignment bottlenecks (S. Figure 1). If supported, methylation calls were only made using reads with a mapping quality higher than 20.

**Simulated Bisulfite Library Comparisons**

A simulation reference genome was created by sampling approximately 2Mb from each chromosome in the human reference genome (hg38) excluding alternative and sex chromosomes. Briefly, 50bp tiles were randomly sampled from a reference chromosome and included in the simulation reference if the tile contained less than 10 ambiguous bases. The first 10kb of the simulated chr1 was duplicated and added as an additional contig. A series of directional and undirectional bisulfite sequencing libraries were then simulated using BSBolt at various read lengths, read depths, and read qualities with random methylation profiles (Table 1). Alignment and methylation calling tools for each package were compared by aligning a simulation library, sorting the alignment file if necessary, and calling methylation values. Each simulation library was processed by each comparison package sequentially in random order on the same compute node. Read alignments were evaluated by the alignment location and strand. An on-target alignment was defined as a read where 95% of the aligned bases were mapped within the simulated region and mapped to the correct origin strand. An alignment was considered off-target if fewer than 5% of the aligned bases were mapped to the simulation region,

the aligned strand of origin was incorrect or flagged as a quality control failure. Accuracy of the CpG methylation calls were evaluated by comparing the called methylation value with the simulated value.

**Targeted Bisulfite Library Comparisons**

We next utilized publicly available targeted bisulfite sequencing data (GSE152923) generated from peripheral blood mononuclear cells of four individuals [25].The libraries were generated using the SureSlectXT Methyl-Seq (Aligent) kit and three sequencing libraries were generated for each individual with varying levels of input DNA (1000ng, 300-1000ng, and 150ng-300ng). Each library was sequenced (100bp, paired end) on an Illumina NovaSeq generating an average of 144.1 million (118.5 - 230.5) paired end reads. In addition to the sequencing data, methylation measurements were generated using the Infinium MethylationEPIC array (Illumina) for all four individuals. Whole genome bisulfite alignment indices were generated using hg38 for each bisulfite sequencing package. Every sequencing library was aligned and processed using the same workflow. Alignment files were generated, duplicate reads were marked using samtools (v1.9), and methylation values were called. Each alignment and methylation calling workflow was given a maximum runtime of 288 hours. If an alignment was incomplete at the end of 288 hours, duplicate read marking and methylation calling was performed on the reads aligned during the 288 hour limit. Methylation calls made for CpG sites with more than five reads covering a site were then compared with array methylation values from the same biological sample.

**Results**

BSBolt was the fastest alignment tool across all simulation conditions, aligning close to 2.29 million reads per minute on average (Figure 2A). BSBolt was approximately 40% faster than the next fastest alignment tool, BISCUIT. When looking at alignment performance by library type, BISCUIT exhibited similar performance to BSBolt when aligning directional reads, but was approximately 229% slower aligning undirectional libraries (Figure 2A). BSSeeker2, BWA-Meth, and Bismark were slower than both BSBolt and BISCUIT when aligning all library types (Figure 2A). BSBolt and BISCUIT aligned the majority of simulated reads across all conditions (>99%) with high accuracy (>99%). BWA-Meth aligned the majority of reads accurately for directional libraries, but as undirectional libraries are unsupported, BWA-Meth undirectional alignments had low mappability (□=0.724) and a low proportion of aligned reads were on target (□=0.706). BSSeeker2 and Bismark exhibited the lowest average mappability across all simulation conditions at 93.6% and 86.9% respectively but the output alignments were generally accurate (Figure 2B). Moreover, BSSeeker2 and Bismark aligned a low percentage of the simulated reads, 65.3% and 42.4% respectively, when the simulated sequencing error and genetic variation was increased from 0.05% to 2% (S. Table 1). Bismark and BSSeeker2 both discard base call quality information when aligning reads so the low mappability with error prone reads is expected.

BSBolt methylation calling was significantly faster than all other tools, with a roughly 11 fold performance advantage over the next fastest tools, BISCUIT and BWA-Meth. BSeeker2 and Bisamark were considerably slower and exhibited a strong relationship between call time and the number of simulated reads (Figure 2C). We also looked at the mean absolute error (MAE) between the number of reads simulated at a given position and the number of reads utilized by each tool to call methylation. BSBolt had the lowest average MAE (0.11 reads) followed by BISCUIT (0.70 reads) and Bismark (0.76 reads). BWA-Meth and BSSeeker2 exhibited high coverage MAE at 6.12 and 8.69 reads respectively. While the BSSeeker2 coverage MAE was high it was not strand biased and the methylation level MAE was small, 0.024. By contrast, the methylation calls made by BWA-Meth were strand biased as shown by the methylation value MAE, 0.255. Overall, BSBolt had the lowest observed methylation level MAE (0.002) followed by BISCUIT (0.013) and Bismark (0.024) (Figure 2D).

The performance of each tool with the targeted bisulfite sequencing libraries largely mirrored the results with the simulation data. However, even though the targeted libraries are directional, BSBolt outperformed BISCUIT aligning an average of 663k reads per minute compared with 637k (Figure 3A). BSSeeker2 failed to align three sequencing libraries within the 288 hour alignment limit, aligning only 78% of reads on average. BSBolt was the fastest methylation calling tool, calling CpG methylation in just 4.35 minutes on average (Figure 3B). We then compared the absolute differences between the sequencing and Illumina EPIC array calls made for the same biological sample, excluding BSSeeker2 alignments as three alignments were incomplete. The absolute differences for all comparisons were combined by tool and binned by effective read coverage, or the number of reads used to call the methylation value (Figure 3C). The called methylation values were highly correlated with the sites called on the EPIC array across all alignment tools (Pearson's r=.92-98, S. Table 2), as previously reported [25]. Unsurprisingly, as sequencing depth increases the observed mean absolute deviation decreases for all tools. At sequencing depths above 40 reads per CpG BSBolt has the smallest absolute deviation between the sequencing and array calls. Note, due the design of the targeted bisulfite libraries, DNA from one origin strand is preferentially captured over a given region. As a result, the strand bias of the BWA-Meth methylation caller didn't noticeably impact the methylation calls.

**Discussion**

Both BSBolt and BISCUIT are significantly faster at bisulfite read alignment while also being more accurate on average than BSSeeker2, Bismark, and BWA-Meth. BSBolt offered marginal performance improvement over BISCUIT with real directional bisulfite libraries, but a large performance gain for the simulated undirectional libraries due to the implementation of a pre-alignment sequencing assessment step. In addition to aligning each read, BSBolt calls contextual read level methylation and assesses read bisulfite conversion, generating alignment information similar to Bismark and BSSeeker2. Importantly, as Bismark and BSSekeer2 have been widely adopted by the community at large it is important to provide

the same alignment information to preserve compatibility with downstream tools. BISCUIT offers support for read bisulfite conversion assessment but it is implemented as post-alignment utility. The BSBolt methylation caller was significantly faster than other tools while also providing more accurate methylation calls. Much of this improvement can be attributed to the structuring read alignment before output; by modifying the alignment strand to reflect the bisulfite origin strand methylation calls can be made rapidly without the need to perform additional formatting.

BSBolt is implemented as a python package installable through the python package index[26] and the Anaconda package manager[27]. In addition to a fully command line interface each BSBolt module can be executed natively as an object in a python (>3.6) environment; providing flexibility for informatics pipelines. BSBolt is available at [28] and is released under the MIT license.

**Availability and requirements**
**Project name**: BSBolt
**Project home pag**e: https://github.com/NuttyLogic/BSBolt
**Operating system(s)**: Platform Independent
**Programming language**: Python >= 3.6
**Other requirements**: numpy>=1.16.3, tqdm>=4.31.1
**License**: MIT
**RRID**: SCR_019080

**Data Availability**
Targeted bisulfite sequencing and EPIC array data deposited in GEO, GSE152923. The pipeline used to simulate bisulfite sequencing libraries is deposited in the analysis repository[28]. Supporting materials and analysis code for this paper are also available in GitHub[29]. Code snapshots and other supporting data are available in the *GigaScience* GigaDB database [30].

**Acknowledgments and Funding**

This work was supported by the National Institutes of Health (T32CA201160 to C.F.). This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

**Supplementary Information**
**Supplemental Table 1**: Simulated Bisulfite Sequencing Library Run Stats
**Supplemental Table 2**: Targeted Bisulfite Alignment Stats
**Supplemental Figure1**: Samtools BAM Conversion Thread Comparisons
**Supplemental Figure2**: BSBolt Performance Characteristics on 150bp Simulated Libraries

**References**


1. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. Science. 2010;328:916–9.

2. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al.. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500:477–81.

3. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology. 2013;14:R115.

4. Orozco LD, Farrell C, Hale C, Rubbi L, Rinaldi A, Civelek M, et al.. Epigenome-wide association in adipose tissue from the METSIM cohort. Hum Mol Genet. 2018;27:2586.

5. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nature Reviews Genetics. 2013;14:204–20.

6. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33:5868–77.

7. Morselli M, Farrell C, Rubbi L, Fehling HL, Henkhaus R, Pellegrini M. Targeted bisulfite sequencing for biomarker discovery. Methods. 2020; doi: 10.1016/j.ymeth.2020.07.006.

8. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al.. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci. 1992;89:1827–31.

9. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al.. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008;452:215–9.

10. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al.. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.

11. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.

12. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, et al.. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics. 2013;14:774.

13. Pedersen BS, Eyring K, De S, Yang IV, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. arXiv 2014;arXiv:1401.1129.

14. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;arXiv: 1303.3997.

15. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

16. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP. Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat Genet. 2017;49:719–29.

17. Guo S, Diep D, Plongthongkum N, Fung H-L, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet. 2017;49:635–42.

18. Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, et al.. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res. 2018;46:e89.

19. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, Keane T, Davies RM. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience. 2021 Feb 16;10(2):giab007. doi: 10.1093/gigascience/giab007.

20. biscuit 2020. bsicuit (Version 0.3.16.20200420); https://github.com/zhou-lab/biscuit

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.

22. pysam 2020. pysam (Version 0.16.1); https://github.com/pysam-developers/pysam

23. Li H. Wgsim 2010. Wgsim (Version 0.3.1-r13); https://github.com/lh3/wgsim

24. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008. doi: 10.1093/gigascience/giab008.

25. Shu C, Zhang X, Aouizerat BE, Xu K. Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells. Epigenetics & Chromatin. 2020; doi: 10.1186/s13072-020-00372-6.

26. PyPI · The Python Package Index. https://pypi.org/. Accessed 11 Sep 2020.

27. Anaconda Software Distribution. Anaconda Documentation. Anaconda Inc. https://anaconda.org; Accessed 11 Sep 2020.

28. BSBolt 2021. BSBolt (Version 1.4.4). https://pypi.org/project/BSBolt/

29. Farrell C; Thompson M; Tosevska A; Oyetunde A; Pellegrini M. Supporting Material and Analysis Code for BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform. GitHub https://github.com/NuttyLogic/BSBoltManuscript

30. Farrell C; Thompson M; Tosevska A; Oyetunde A; Pellegrini M. Supporting data for "BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform" GigaScience Database 2021. http://dx.doi.org/10.5524/100879

**Table 1: Simulated Bisulfite Sequencing Library Parameters:** The parameters used with BSBolt Simulate to prepare simulation libraries using BSBolt for tool comparisons.

| Read Depth | Mutation Rate | Sequencing Error | Sequencing Type | Library Type |
|---|---|---|---|---|
| 20 | 0.005 | 0.005 | Paired End | Directional |

| | | | | |
|---|---|---|---|---|
| 20 | 0.005 | 0.005 | Paired End | Directional |
| 20 | 0.005 | 0.005 | Paired End | Directional |
| 30 | 0.005 | 0.005 | Paired End | Undirectional |
| 30 | 0.005 | 0.005 | Paired End | Undirectional |
| 30 | 0.005 | 0.005 | Paired End | Undirectional |
| 20 | 0.005 | 0.005 | Single End | Directional |
| 20 | 0.005 | 0.005 | Single End | Directional |
| 20 | 0.005 | 0.005 | Single End | Directional |
| 30 | 0.005 | 0.005 | Single End | Undirectional |
| 30 | 0.005 | 0.005 | Single End | Undirectional |
| 30 | 0.005 | 0.005 | Single End | Undirectional |
| 8 | 0.005 | 0.005 | Paired End | Directional |
| 8 | 0.005 | 0.005 | Paired End | Directional |
| 8 | 0.005 | 0.005 | Paired End | Directional |
| 8 | 0.005 | 0.005 | Single End | Directional |
| 8 | 0.005 | 0.005 | Single End | Directional |
| 8 | 0.005 | 0.005 | Single End | Directional |
| 8 | 0.01 | 0.02 | Paired End | Directional |
| 8 | 0.01 | 0.02 | Paired End | Directional |
| 8 | 0.01 | 0.02 | Paired End | Directional |

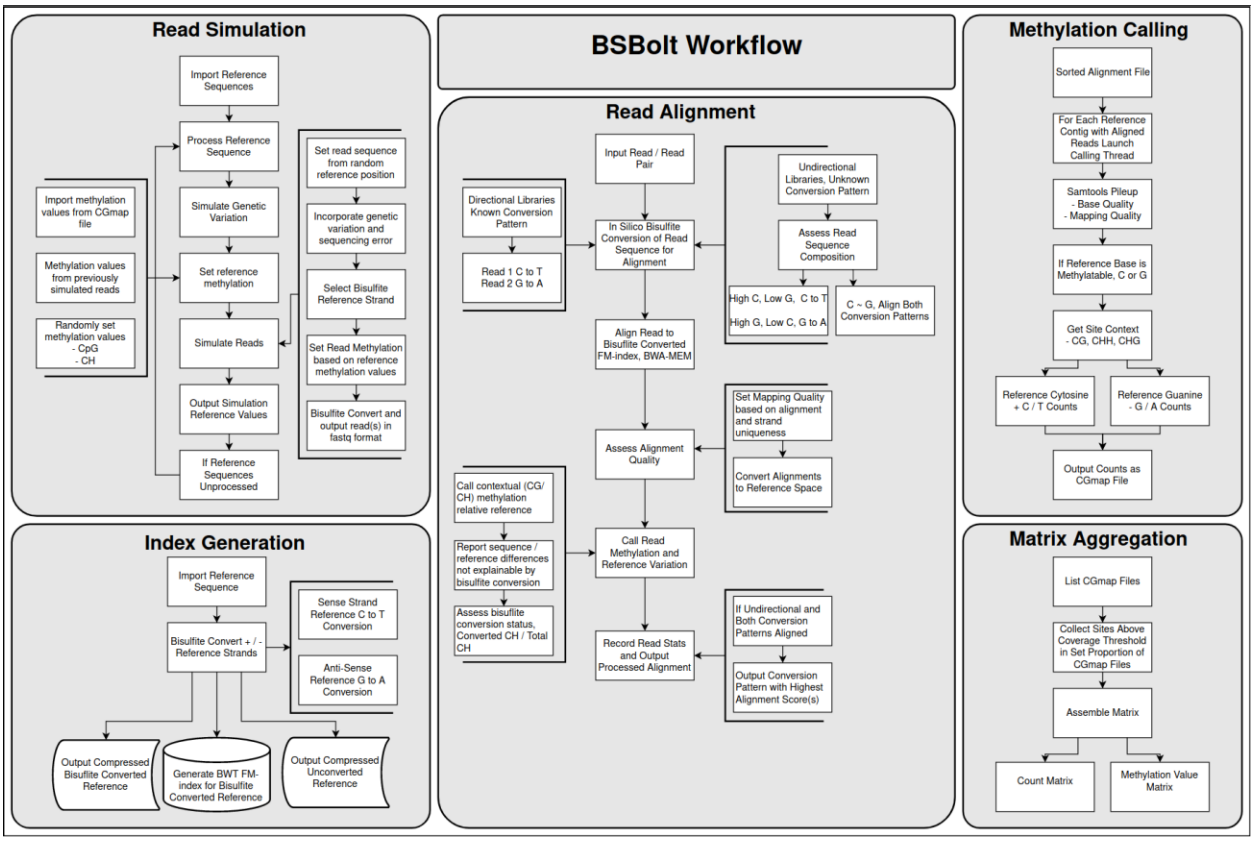



**Figure1: BSBolt Workflows**

BSBolt is implemented as a series of discrete modules for read simulation, index generation, read alignment, methylation calling, and matrix aggregation. All BSBolt modules can be run using a command line interface or within a python (>3.6) environment natively.

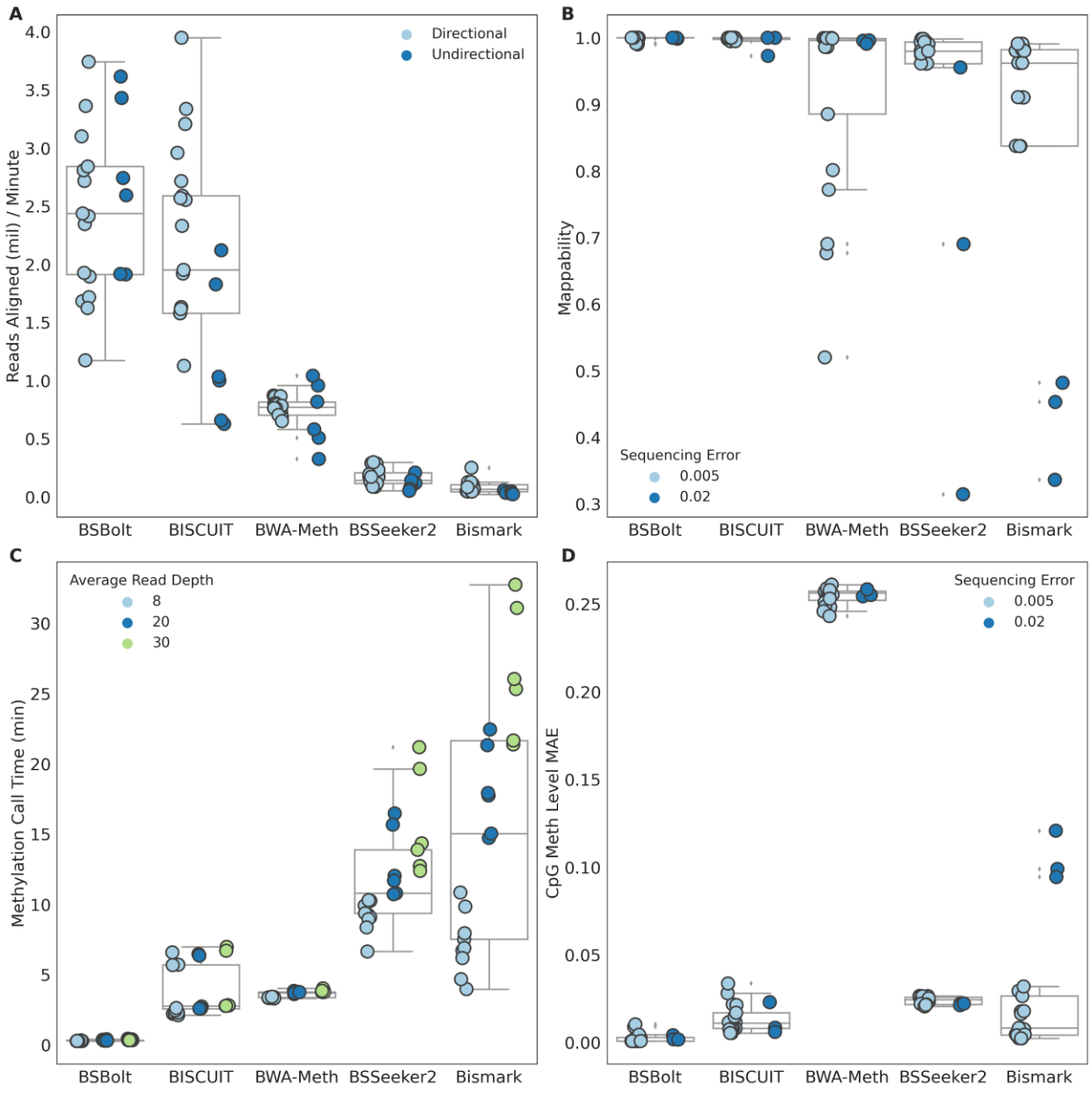



**Figure 2: Simulated Bisulfite Sequencing Library Performance**

(A) Reads aligned per minute for each bisulfite alignment tool. (B) Proportion of simulated reads mapped during alignments. Note, BWA-Meth does not support undirectional library alignment resulting in low mappability for undirectional libraries. (C) Methylation call time (min) for each alignment tool. (D) Mean Absolute Error (MAE) observed between the simulated and called methylation value.

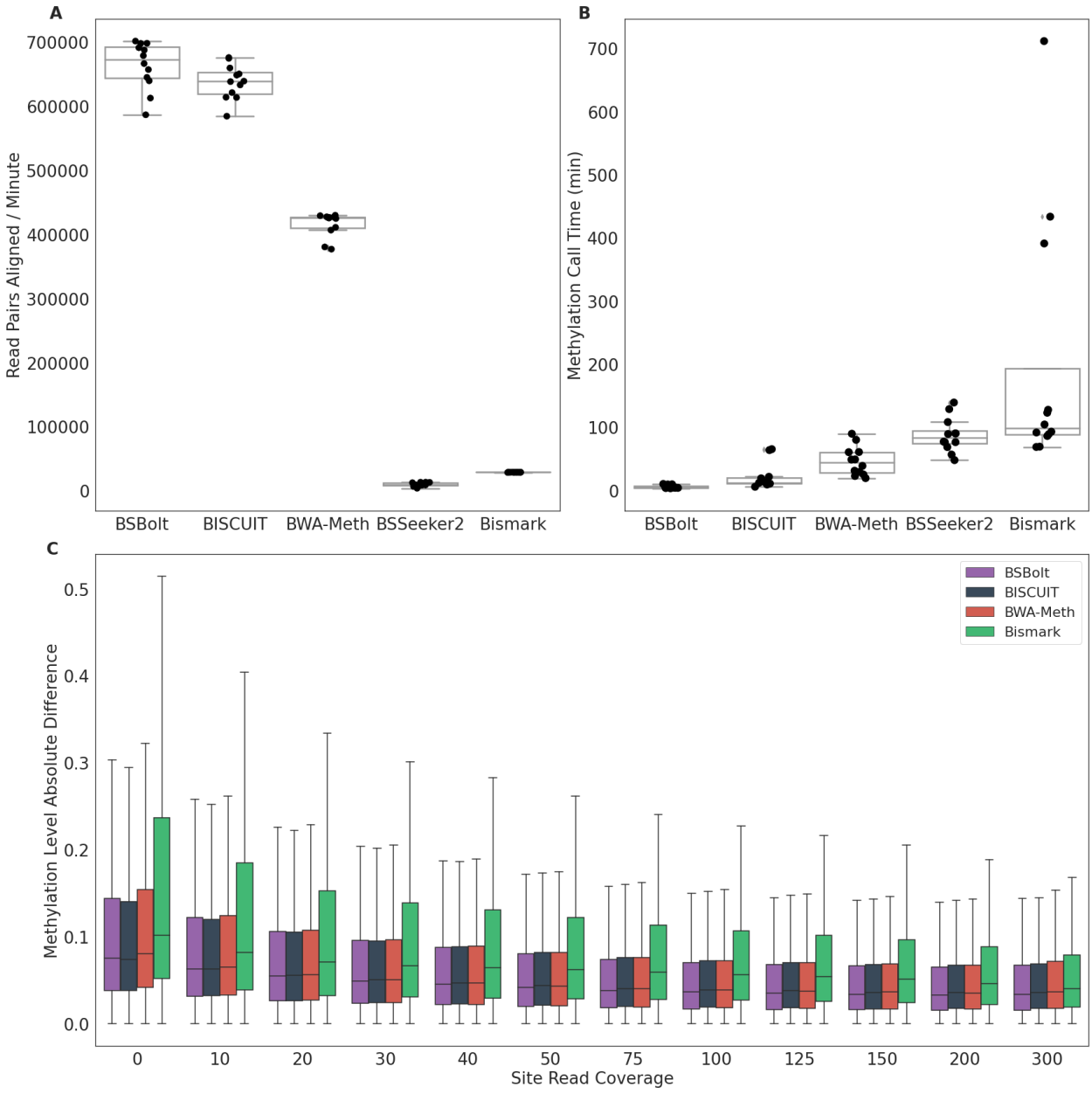



**Figure 3: Targeted Bisulfite Sequencing Library Performance**
(A) The number of read pairs aligned per minute for each bisulfite alignment tool. (B) Total methylation calling time (min) for each alignment file. (C) The absolute difference between array methylation values and sequencing methylation values for overlapping calls, binned by effective read depth.

**Supplemental Figure1:**

Alignment times for 150 base pair simulated libraries by the number of threads used for SAM to BAM conversion.

**Supplemental Figure 2:**

(A) Total alignment time (min) and (B) Maximum memory utilization (mb) for simulated 150 bp bisulfite sequencing libraries by the number of alignment threads provided to BSBolt.

Table 1 Click here to access/download;Table;BSBolt Table1.xlsx

**Table 1: Simulated Bisulfite Sequencing Library Parameters:** The parameters used to simulate libraries u

| Average Read Depth | Mutation Rate | Sequencing Error | Sequencing Type | Library Type |
|---|---|---|---|---|
| 20 | 0.005 | 0.005 | Paired End | Directional |
| 20 | 0.005 | 0.005 | Paired End | Directional |
| 20 | 0.005 | 0.005 | Paired End | Directional |
| 30 | 0.005 | 0.005 | Paired End | Undirectional |
| 30 | 0.005 | 0.005 | Paired End | Undirectional |
| 30 | 0.005 | 0.005 | Paired End | Undirectional |
| 20 | 0.005 | 0.005 | Single End | Directional |
| 20 | 0.005 | 0.005 | Single End | Directional |
| 20 | 0.005 | 0.005 | Single End | Directional |
| 30 | 0.005 | 0.005 | Single End | Undirectional |
| 30 | 0.005 | 0.005 | Single End | Undirectional |
| 30 | 0.005 | 0.005 | Single End | Undirectional |
| 8 | 0.005 | 0.005 | Paired End | Directional |
| 8 | 0.005 | 0.005 | Paired End | Directional |
| 8 | 0.005 | 0.005 | Paired End | Directional |
| 8 | 0.005 | 0.005 | Single End | Directional |
| 8 | 0.005 | 0.005 | Single End | Directional |
| 8 | 0.005 | 0.005 | Single End | Directional |
| 8 | 0.01 | 0.02 | Paired End | Directional |
| 8 | 0.01 | 0.02 | Paired End | Directional |
| 8 | 0.01 | 0.02 | Paired End | Directional |

using BSBolt for tool comparisons

Figure1

Click here to access/download;Figure;Figure1.png

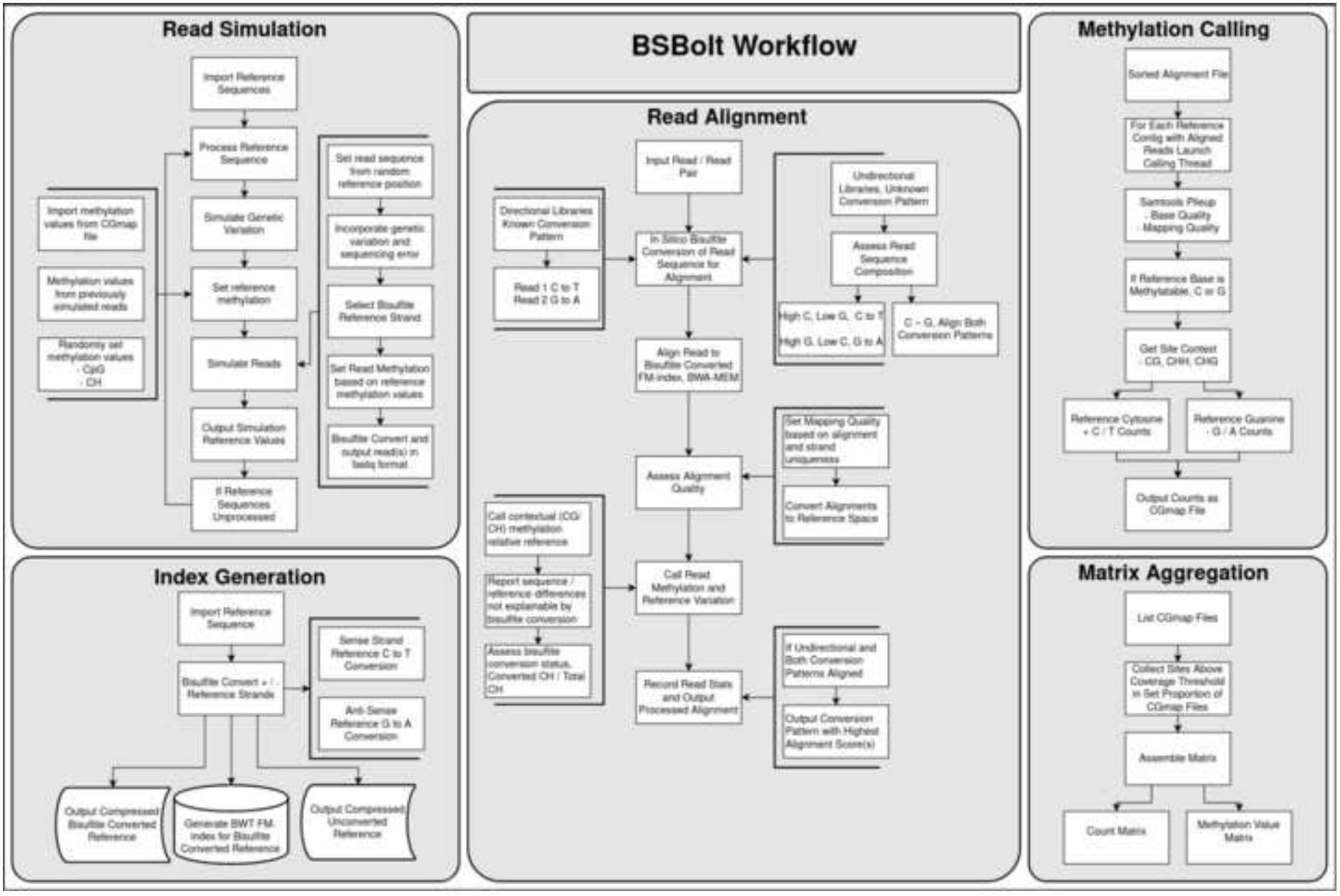


Figure2



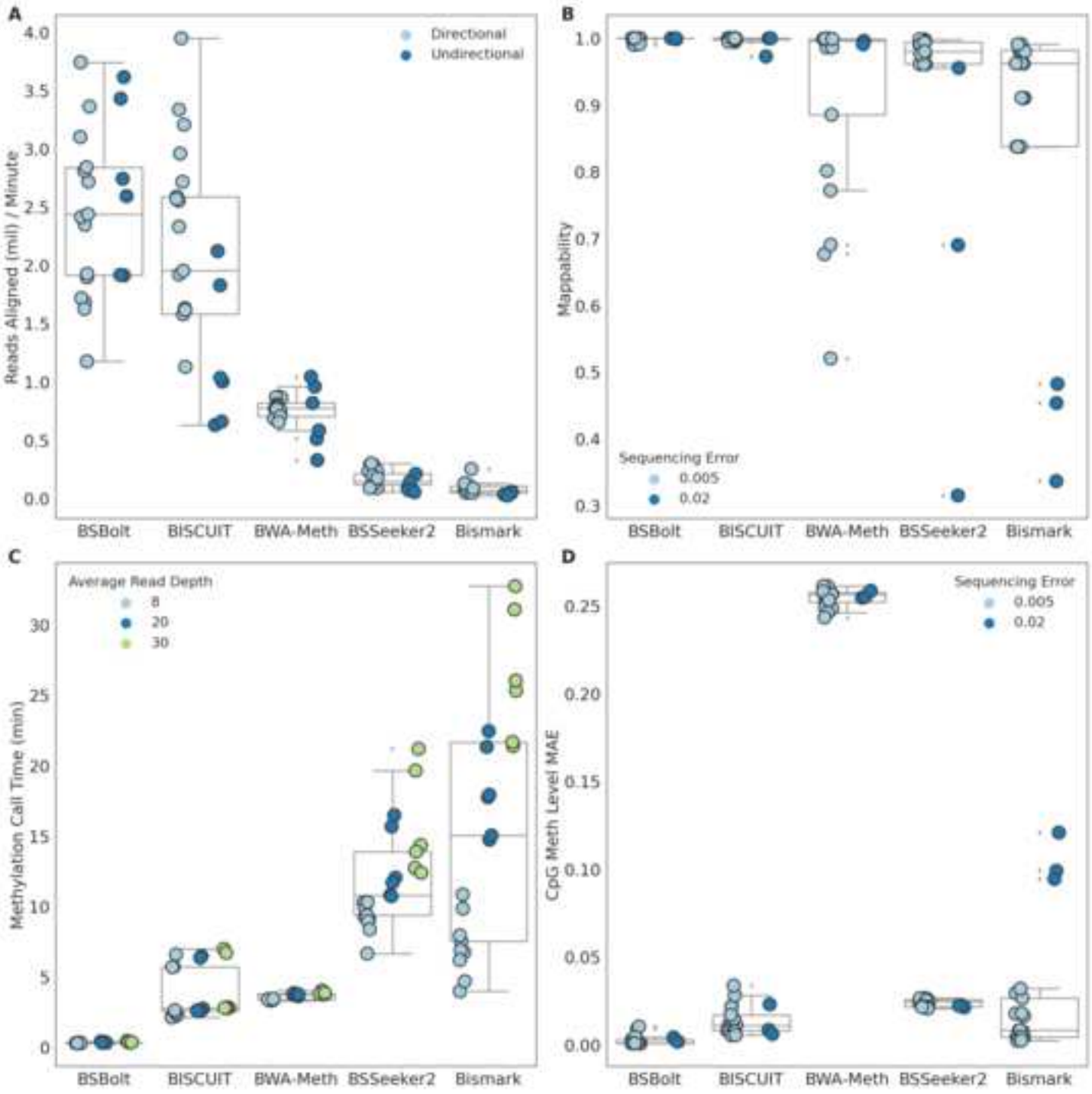
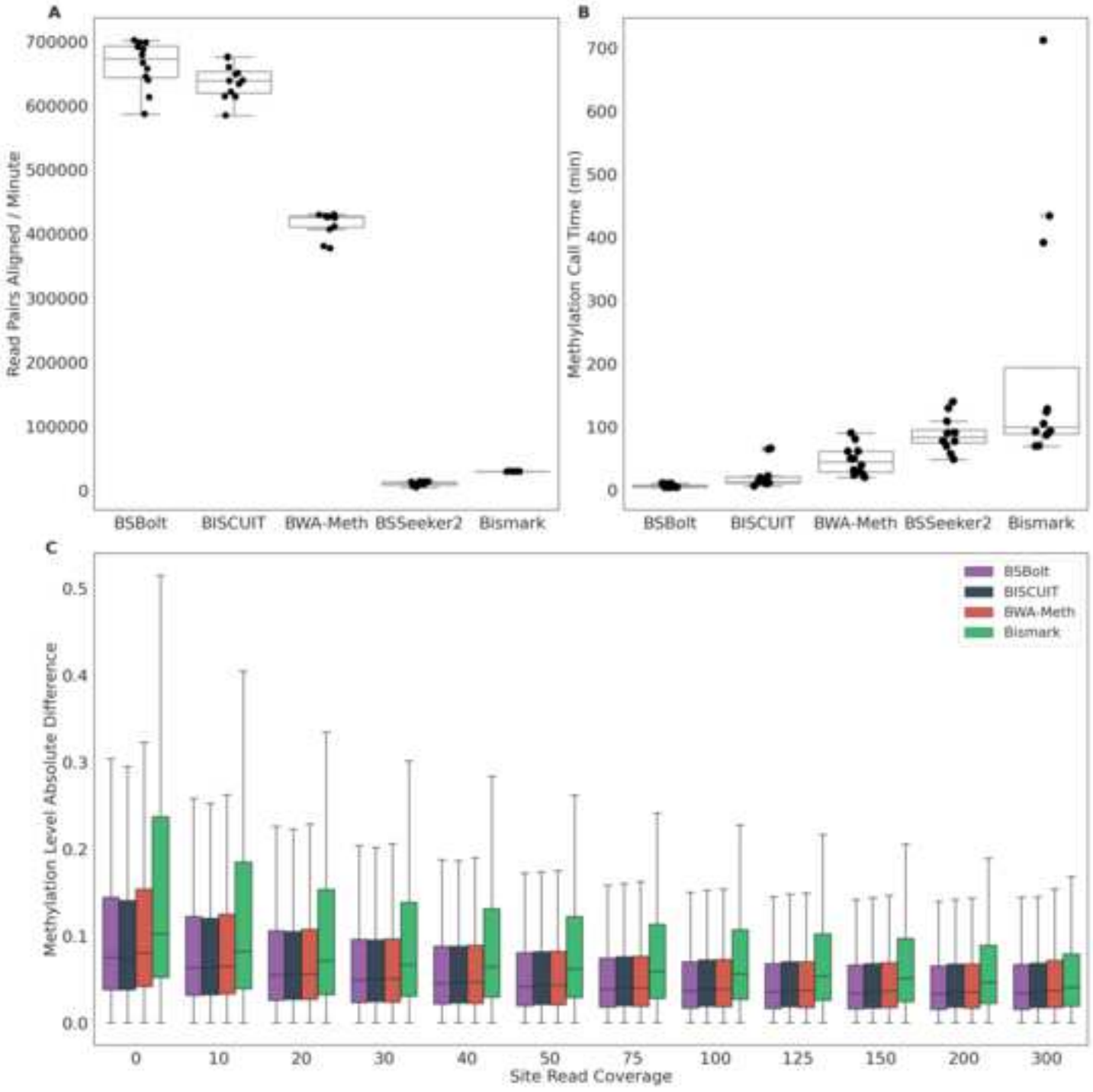


Figure3

Click here to access/download;Figure;Figure3.png



Supplementary Table 1

Click here to access/download

**Supplementary Material**

BSBolt Supplemental Table 1.xlsx

Supplementary Table 2

Click here to access/download

**Supplementary Material**

BSBolt Supplemental Table 2.xlsx

Supplementary Figure 1

Supplementary Figure 2