BiSulfite Bolt: A BiSulfite Sequencing Analysis Platform

Colin Farrell[1], Michael Thompson[2], Anela Tosevska[2], Adewale Oyetunde[2], Matteo Pellegrini[2,3]

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;
3 Corresponding Author;

Corresponding Author
Matteo Pellegrini
matteop@mcdb.ucla.edu

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.
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 a 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(Zemach et al., 2010; Ziller et al., 2013) and numerous biological processes (Horvath, 2013; Orozco et al., 2018; Smith & Meissner, 2013). High throughput sequencing combined with bisulfite conversion is a broadly used method for profiling DNA methylation genome wide(Meissner et al., 2005) (Morselli et al., 2020). 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 (Frommer et al., 1992). 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 (Cokus et al., 2008; Frommer et al., 1992; Lister et al., 2009). 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 such as Bismark(Krueger & Andrews, 2011), BS-Seeker2(W. Guo et al., 2013; Krueger & Andrews, 2011), and BWA-Meth(Pedersen et al., 2014) successfully adopted a three-base alignment strategy wrapped around established read aligners such as Bowtie2(Langmead & Salzberg, 2012; H. Li, 2013) and BWA-MEM(H. Li, 2013), 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 (S. Guo et al., 2017; Jenkinson et al., 2017; W. Li et al., 2018). 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. While each of these tools is capable of outputting accurate bisulfite read alignments, wrapping external read alignment tools introduces added complexity which can negatively impact alignment performance and in turn methylation assessment.

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. BSBolt alignment is built on a forked version BWA-MEM(H. Li, 2013; H. Li et al., 2009) and HTSLIB(H. Li et al., 2009) with bisulfite specific sequencing logic integrated directly into the alignment process. 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(*Biscuit*, n.d.)**.** 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 a robust read simulation utility and a tool for 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. The 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. In this case BSBolt first analyzes the read base composition. A read, or read pair, with a low proportion observed cytosines compared to guanine 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(H. Li et al., 2009), and initialized using pysam(*Pysam*, n.d.), 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(H. Li, n.d.) 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. 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 approximatley 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 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 (Shu et al., n.d.).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 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. 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 (Table 2). BSBolt was approximately 30% faster than the next fastest alignment tool, BISCUIT. When looking at alignment performance by library type, BISCUIT was approximately 8% faster than BSBolt when aligning directional reads, but approximately 229% slower aligning undirectional libraries (Table 2). BSSeeker2, BWA-Meth, and Bismark were slower than both BSBolt and BISCUIT when aligning all library types (Table 2). 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 (μ=0.724) and a low proportion of aligned reads were on target (μ=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 (Table 2). 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 (S. Table 1). 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).

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 653k reads per minute compared with 633k (Figure 2A). Neither Bismark nor BSSeeker2 aligned any of the sequencing libraries within the 24 hour alignment limit, aligning 29.11% and 12.9% of the read pairs respectively. Even though the alignment files for Bismark and BSSeeker2 were considerably smaller than the other alignment tools, methylation calling by the other packages was faster, with BSBolt calling CpG methylation in just 4.35 minutes on average (Figure 2B). We then compared the absolute differences between the sequencing and Illumina EPIC array calls made for the same biological sample. 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 2C). BSSeeker2 was excluded from this analysis due to few overlapping sequencing and array methylation calls. 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. 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(*PyPI · The Python Package Index*, n.d.). This streamlines the installation process for newer users. During the installation process a pre-compiled system specific binary is automatically installed, or compiled automatically if a system binary is unavailable. In addition to a fully command line interface each BSBolt module can be executed natively as an object in a python (>3.5) environment; providing flexibility for informatics pipelines. BSBolt is available at https://pypi.org/project/BSBolt/ and is released under the MIT license.

Availability and requirements

Project name : BSBolt

Project home page : 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

Acknowledgments and Funding

This work was supported by theNational Institutes of Health (T32CA201160 to C.F.).

Supplemental Information

Analysis Code: nuttylogic.github.com/BSBoltManuscript

Supplemental Table 1: Simulated Bisulfite Sequencing Library Run Stats

Supplemental Table 2: Average Targeted Bisulfite Alignment Stats

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.

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.

biscuit. (n.d.). Github. Retrieved September 8, 2020, from https://github.com/zhou-lab/biscuit

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., & Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, *452*(7184), 215–219.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., & Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(5), 1827–1831.

Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K., & Zhang, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nature Genetics*, *49*(4), 635–642.

Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M. Q., Chen, P.-Y., & Pellegrini, M. (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, *14*, 774.

295 Horvath, S. (2013). DNA methylation age of human tissues and cell types. In *Genome Biology*
296 (Vol. 14, Issue 10, p. R115).
297 Jenkinson, G., Pujadas, E., Goutsias, J., & Feinberg, A. P. (2017). Potential energy landscapes
298 identify the information-theoretic nature of the epigenome. *Nature Genetics*, *49*(5), 719–
299 729.
300 Krueger, F., & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for
301 Bisulfite-Seq applications. In *Bioinformatics* (Vol. 27, Issue 11, pp. 1571–1572).
302 https://doi.org/10.1093/bioinformatics/btr167
303 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature*
304 *Methods*, *9*(4), 357–359.
305 Li, H. (n.d.). *wgsim*. Github. Retrieved September 8, 2020, from https://github.com/lh3/wgsim
306 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
307 In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997
308 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,
309 Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence
310 Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.
311 Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R.,
312 Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V.,
313 Millar, A. H., Thomson, J. A., Ren, B., & Ecker, J. R. (2009). Human DNA methylomes at
314 base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322.
315 Li, W., Li, Q., Kang, S., Same, M., Zhou, Y., Sun, C., Liu, C.-C., Matsuoka, L., Sher, L., Wong,
316 W. H., Alber, F., & Zhou, X. J. (2018). CancerDetector: ultrasensitive and non-invasive
317 cancer detection at the resolution of individual reads using cell-free DNA methylation
318 sequencing data. *Nucleic Acids Research*, *46*(15), e89.
319 Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005).
320 Reduced representation bisulfite sequencing for comparative high-resolution DNA

methylation analysis. *Nucleic Acids Research*, *33*(18), 5868–5877.

Morselli, M., Farrell, C., Rubbi, L., Fehling, H. L., Henkhaus, R., & Pellegrini, M. (2020). Targeted bisulfite sequencing for biomarker discovery. *Methods* . https://doi.org/10.1016/j.ymeth.2020.07.006

Orozco, L. D., Farrell, C., Hale, C., Rubbi, L., Rinaldi, A., Civelek, M., Pan, C., Lam, L., Montoya, D., Edillor, C., Seldin, M., Boehnke, M., Mohlke, K. L., Jacobsen, S., Kuusisto, J., Laakso, M., Lusis, A. J., & Pellegrini, M. (2018). Epigenome-wide association in adipose tissue from the METSIM cohort. *Human Molecular Genetics*, *27*(14), 2586.

Pedersen, B. S., Eyring, K., De, S., Yang, I. V., & Schwartz, D. A. (2014). Fast and accurate alignment of long bisulfite-seq reads. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1401.1129

PyPI · The Python Package Index. (n.d.). Retrieved September 11, 2020, from https://pypi.org/

pysam. (n.d.). Github. Retrieved September 8, 2020, from https://github.com/pysam-developers/pysam

Shu, C., Zhang, X., Aouizerat, B. E., & Xu, K. (n.d.). *Comparison of Methylation Capture Sequencing and Infinium EPIC Methylation Array in Peripheral Blood Mononuclear Cells*. https://doi.org/10.21203/rs.3.rs-33940/v1

Smith, Z. D., & Meissner, A. (2013). DNA methylation: roles in mammalian development. In *Nature Reviews Genetics* (Vol. 14, Issue 3, pp. 204–220). https://doi.org/10.1038/nrg3354

Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D. (2010). Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. In *Science* (Vol. 328, Issue 5980, pp. 916–919). https://doi.org/10.1126/science.1186366

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., & Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, *500*(7463), 477–481.

Table 1: Simulated Bisulfite Sequencing Library Parameters

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

Table 2: Simulated Bisulfite Sequencing Average Run Performance

Tool (Library Type)	Mappability (%)	On Target / Tot. Align. (%)	Off Target / Tot. Align. (%)	Alignment Time (min)	Mil. Reads Aligned / Min.	CpG Meth Level MAE	CpG Meth Level STD	CpG Coverage MAE	CpG Coverage STD	Meth. Call Time (min)	Comparison Libraries
BWA-Meth (Undirectional)	72.44%	70.64%	29.36%	22.784	0.705	0.258	0.204	6.177	3.912	3.840	6
BWA-Meth (Directional)	99.63%	99.88%	0.12%	8.612	0.773	0.253	0.225	6.102	2.489	3.513	15
BWA-Meth (All Libraries)	91.86%	91.53%	8.47%	12.661	0.754	0.255	0.219	6.124	2.895	3.607	21
BISCUIT (Undirectional)	99.89%	99.79%	0.21%	13.373	1.212	0.016	0.030	1.246	1.284	4.145	6
BISCUIT (Directional)	99.72%	99.73%	0.27%	2.663	2.403	0.012	0.033	0.487	0.693	3.682	15
BISCUIT (All Libraries)	99.77%	99.75%	0.25%	5.723	2.063	0.013	0.032	0.704	0.862	3.814	21
BSBolt (Undirectional)	99.83%	99.72%	0.28%	6.460	2.428	0.003	0.018	0.203	0.573	0.362	6
BSBolt (Directional)	99.87%	99.77%	0.23%	2.872	2.242	0.002	0.020	0.066	0.257	0.307	15
BSBolt (All Libraries)	99.86%	99.76%	0.24%	3.897	2.295	0.002	0.020	0.105	0.347	0.323	21
BSSeeker2 (Undirectional)	98.30%	74.99%	25.01%	145.877	0.114	0.026	0.111	14.734	3.841	15.699	6
BSSeeker2 (Directional)	91.73%	99.98%	0.02%	38.684	0.182	0.023	0.106	6.273	2.441	10.636	15
BSSeeker2 (All Libraries)	93.61%	92.84%	7.16%	69.311	0.162	0.024	0.107	8.691	2.841	12.082	21
Bismark (Undirectional)	94.41%	74.98%	25.02%	425.827	0.036	0.010	0.029	0.822	1.451	26.380	6
Bismark (Directional)	84.00%	100.00%	0.00%	81.112	0.093	0.030	0.069	0.728	0.919	11.589	15
Bismark (All Libraries)	86.97%	92.85%	7.15%	179.602	0.077	0.024	0.057	0.755	1.071	15.815	21

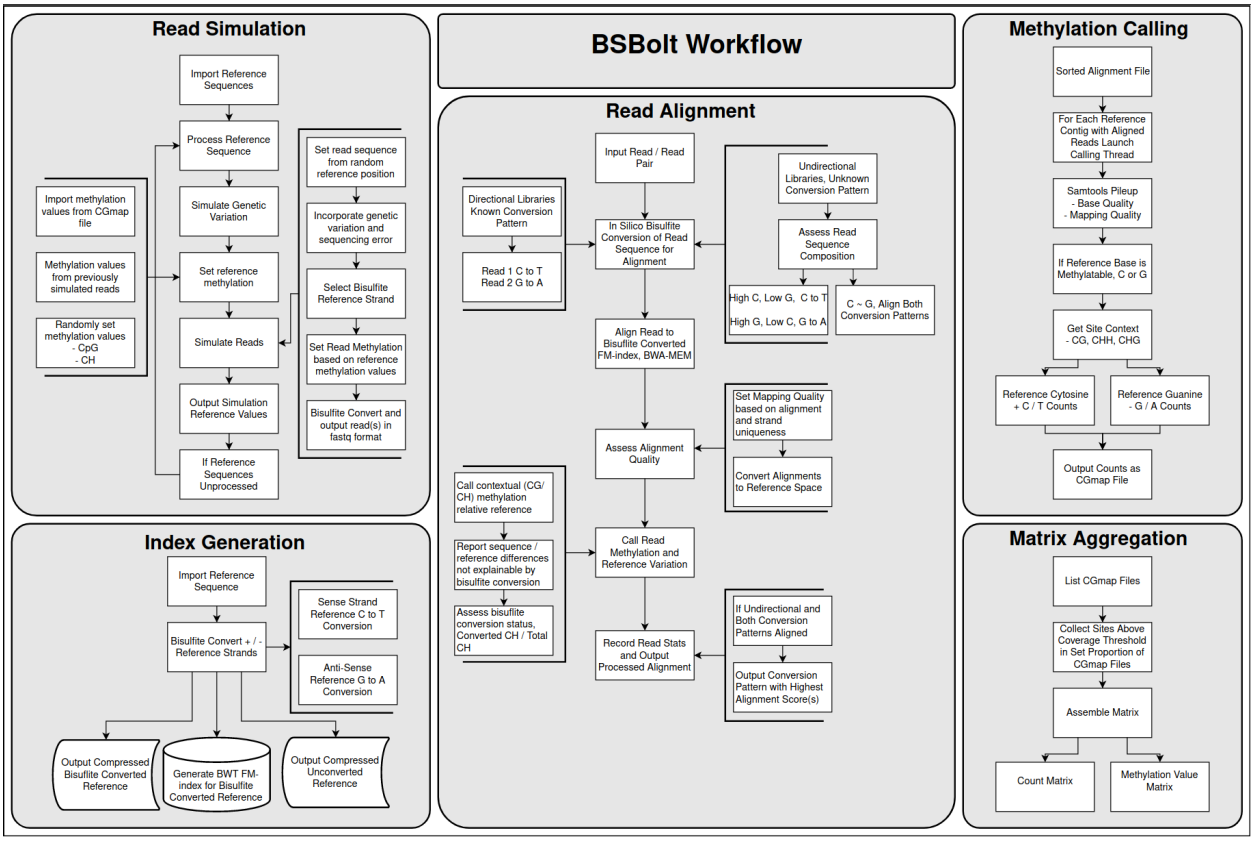

Figure1: BSBolt Workflows

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 (>3.5) environment natively.

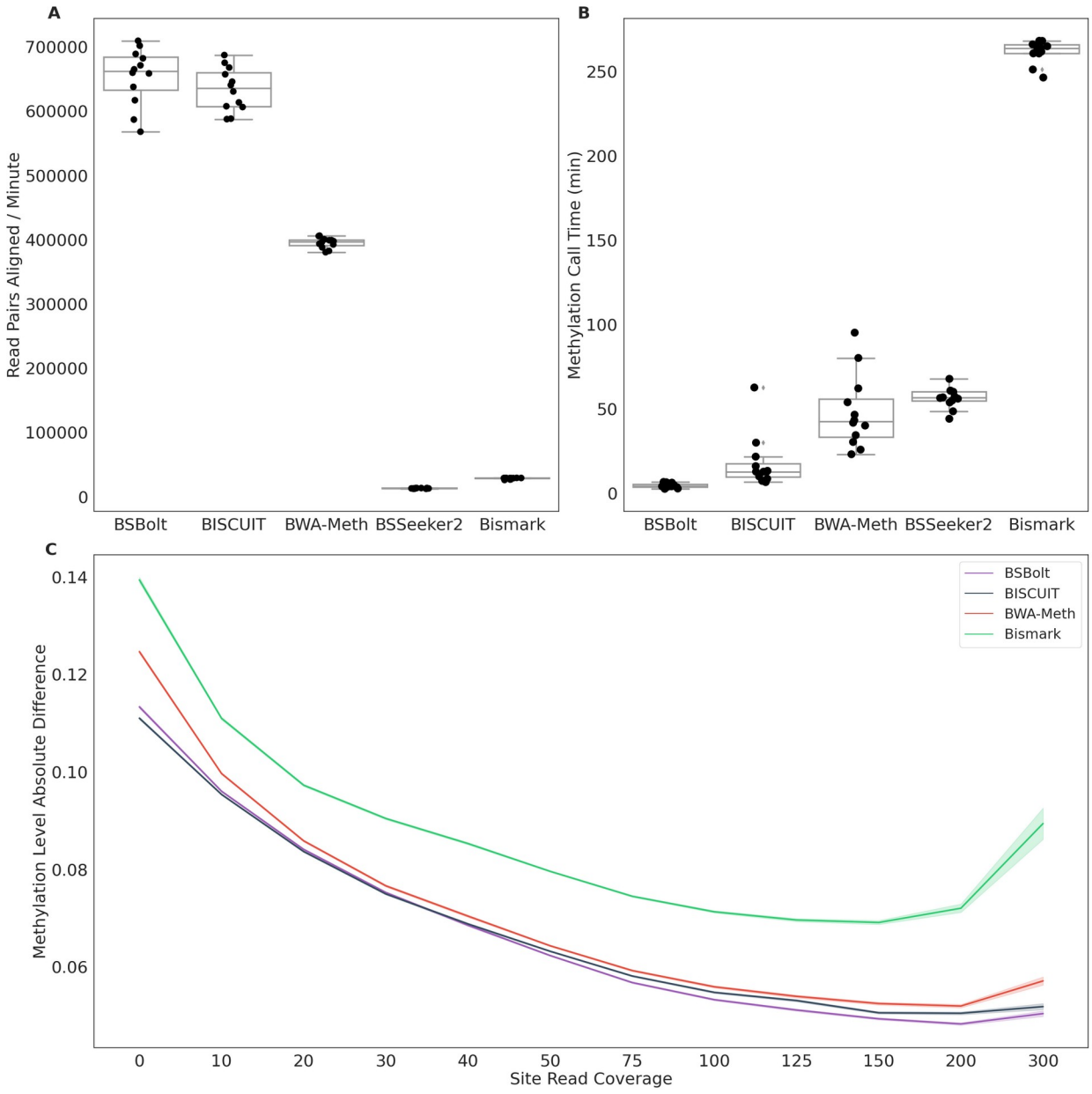

Figure 2: Targeted Bisulfite Sequencing Library Performance
(A) The number of read pairs aligned per minute for each bisulfite alignment tool. (B) Total methylation calling time (min) for each alignment file. (C) The absolute difference between array methylation values and sequencing methylation values for overlapping calls binned by effective read depth. The fit line represents the mean absolute difference at each read depth with a shaded 95% confidence interval computed by bootstrapping (n=10).